Rapid Method Development in Ion Chromatography using Quantitative Structure-Retention Relationships

by

Soo Hyun Park

A dissertation submitted in fulfilment of the requirements

for the degree of

Doctor of Philosophy

School of Physical Sciences

(Chemical Sciences)

University of Tasmania

April 2017

Declaration of Originality

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

19 April 2017

Authority of Access

This thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

19 April 2017

Statement regarding published work contained in this thesis

The publishers of the papers comprising Chapters 3 to 5 hold the copyright for that content, and access to the material should be sought from the respective journals. The remaining non published content of the thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

19 April 2017

Statement of co-authorship

The following people and institutions contributed to the publication of the work undertaken as part of this thesis:

- Candidate: S.H. Park, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- P.R. Haddad, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- R.I.J. Amos, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- M. Talebi, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- M. Taraji, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- Y. Wen, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- R.A. Shellie, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- G.W. Dicinoski, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- G. Schuster, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- E. Tyteca, Australian Centre for Research on Separation Science (ACROSS), School of Physical Sciences, University of Tasmania
- R. Szucs, Pfizer Global Research and Development, Sandwich, UK
- C.A. Pohl, Thermo Fisher Scientific, Sunnyvale, CA, USA
- J.W. Dolan, LC Resources Inc., McMinnville, OR, USA

Author details and their roles:

Paper 1, "Enhanced methodology for porting ion chromatography retention data", Located in Chapter 3

S.H. Park (55%), R.A. Shellie (6%), G.W. Dicinoski (1%), G. Schuster (3%), M. Talebi (5%), P.R. Haddad (20%), R. Szucs (3%), J.W. Dolan (3%), C.A. Pohl (4%)

- S.H. Park was the primary author and with P.R. Haddad and R.A. Shellie contributed to the idea, its formulation and development.
- P.R. Haddad and G. Schuster assisted with refinement and presentation.
- R.A. Shellie and M. Talebi offered general laboratory assistance.
- C.A. Pohl, R. Szucs, J.W. Dolan, and G.W. Dicinoski established the need of study and provided feedback on the work.

Paper 2, "Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model", Located in Chapter 4

S.H. Park (50%), P.R. Haddad (20%), M. Talebi (7%), E. Tyteca (4%), R.I.J. Amos (7%), R. Szucs (6%), J.W. Dolan (3%), C.A. Pohl (3%)

- S.H. Park was the primary author and with P.R. Haddad, M. Talebi, and R.I.J. Amos contributed to the idea, its formulation and development.
- P.R. Haddad, E. Tyteca, and R. Szucs assisted with refinement and presentation.
- R. Szucs offered in-house software package for the modelling.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

Paper 3, "Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. II Use of Tanimoto similarity index in ion chromatography", Located in Chapter 5

S.H. Park (50%), P.R. Haddad (15%), M. Talebi (10%), R.I.J. Amos (8%), E. Tyteca (7%), R. Szucs (4%), J.W. Dolan (3%), C.A. Pohl (3%)

- S.H. Park was the primary author and with P.R. Haddad, M. Talebi, and R.I.J. Amos contributed to the idea, its formulation and development.
- P.R. Haddad, M. Talebi, R.I.J. Amos, and E. Tyteca assisted with refinement and presentation.
- M. Talebi provided general laboratory assistance.
- M. Talebi and E. Tyteca provided a GA-PLS algorithm for the modelling.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

Paper 4, " Towards a chromatographic similarity index to establish localised quantitative structure-retention models for retention prediction: use of retention factor ratio ", Located in Chapter 6

E. Tyteca (50%), M. Talebi (7%), R.I.J. Amos (7%), S.H. Park (5%), M. Taraji (5%), Y. Wen (5%), R. Szucs (2%), C.A. Pohl (2%), J.W. Dolan (2%), P.R. Haddad (15%),

- E. Tyteca was the primary author and with P.R. Haddad, M. Talebi, and R.I.J. Amos contributed to the idea, its formulation and development.
- E. Tyteca, P.R. Haddad, M. Talebi and R.I.J. Amos assisted with refinement and presentation.
- S.H. Park, M. Taraji and Y. Wen offered retention data and molecular descriptors for the modelling, along with preliminary results for pilot study, and offered some descriptions in experimental section of the manuscript.
- R. Szucs, C.A. Pohl and J.W. Dolan established the need of study and provided feedback on the work.

We the undersigned agree with the above stated "proportion of work undertaken" for each of the above published (or submitted) peer-reviewed manuscripts contributing to this thesis:

Signed: _____ _____

Prof. Paul Haddad	*Prof. John Dickey*
Supervisor	*Head of School*
School of Physical Sciences	*School of Physical Sciences*
University of Tasmania	University of Tasmania
Date: 19/4/17	20/4/17

Acknowledgements

First of all, I would like to offer this thesis to God, the almighty, merciful, and passionate, as I do believe that "*in everything God works for the good of those who love him, whom he has called according to his plan* (Romans 8:28)". Especially, I would like to express my deepest appreciation and gratitude to my supervisor Prof. Paul Haddad who gave me this wonderful opportunity to be his student and guided me in the right direction in terms of both research and life. This thesis would not have been possible without his kind support and patience.

Additionally, I am very grateful to the following people:

- My other supervisors Dr. Ruth Amos, Dr. Mohammad Talebi, and Assoc. Prof. Robert Shellie for their kind guidance, assistance and advice during the course of this project.
- Dr. Georg Schuster and Dr. Eva Tyteca for helpful discussions and for their kind assistance in my manuscripts.
- Dr. Phil Zakaria and Dr. Cameron Jones for their kind assistance in basic IC experiments and retention modelling.
- Dr. Roman Szucs for his helpful discussions during his visits, for his interest in my research and for his kind help in proof-reading my manuscripts.
- Mr. Chris Pohl and Dr. John Dolan for helpful discussions and their kind help in proof-reading my manuscripts.
- Ms. Maryam Taraji and Mr. Yabin Wen for their friendship, discussions, and help.
- Past and present members of ACROSS for their friendship and help throughout my PhD course, including Assoc. Prof. Greg Dicinoski, Assoc. Prof. Lito Quirino, Prof. Michael Breadmore, Prof. Brett Paull, Prof. Pavel Nesterenko, Dr. Peter Smejkal, Dr. Alain Wuethrich, Ms. Marina Lanz, Dr. Estrella Rodriguez, Dr. Hong Heng See, Dr.

Sara Sandron, Dr. Naama Karu, Dr. Heide Rabanes, Dr. Aliaa Shallan, Dr. Sinead Currivan, Ms. Marni Tubaon, Ms. Pavisara Nanthasurasak, Ms. Sui Ching Phung, and Ms. Sidra Waheed.

- The staff and students of the School of Physical Sciences (Chemistry Sciences) for providing a friendly and comfortable working environment.
- Prof. Seung Wook Kim, Prof. Byeong Ho Kang, Dr. Hyun-Seob Song, Dr. Sang Jin Moon and Dr. SangBaek Shin for their kind advice.
- People in church for their kind prayers including Fr. Tate, Fr. Brian, Ellen and Terry.

This project was supported by a ARC Linkage Grant (LP120200700) and an International Postgraduate Research Scholarship (IPRS) provided by the Australian Commonwealth Government, as well as a travel grant to attend a overseas conference provided by the UTAS Graduate Research Office. I am very grateful for this funding.

Finally, I would like to say thanks so much to my family for their unconditional love, prayer, encouragement and understanding, which were a big driving force for me to continue and finish this study.

List of abbreviations

ACN	Acetonitrile
APE	Absolute values of the percentage error
CD	Conductivity detection
CPM	Current porting method
CR-TC	Continuously regenerated trap column
CV	Cross validation
GA	Genetic algorithm
GC	Gas chromatography
GDGP	Gradient data for gradient prediction
EA	Evolutionary algorithm
EG	Eluent generator
EGC	EluGen® cartridge
ESI	Electrospary ionisation
HILIC	Hydrophilic-interaction liquid chromatography
IC	Ion chromatography
IDGP	Isocratic data for gradient prediction
IEC	Ion-exchange chromatography
LC	Liquid chromatography
LSS	Linear solvent strength
LSS-EA	LSS model-empirical approach
LV	Latent variable
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MLR	Multiple linear regression
MPM	Modified porting method
MS	Mass spectrometry
MSA	Methanesulfonic acid
NC	Neighbour count
PLS	Partial least squares
QSAR	Quantitative structure-activity relationship
QSRR	Quantitative structure-retention relationship

RMSE	Root-mean-square error
RMSEP	Root-mean-square error of prediction
RPLC	Reversed-phase liquid chromatography
RSD	Relative standard deviation
SRD	Sum of ranking difference
TS	Tanimoto similarity
UV	Ultra-violet
VIF	Variance inflation factor

List of symbols

ΔA	UV absorbance change on elution of a sample
α	Selectivity
C_i	Initial concentration for the gradient elution
C_s	Sample concentration
ε	Molar absorptivity
F	Fisher ratio
F_v	Eluent flow-rate
m	Path-length of the UV detector cell
n_{tot}	A total number of bits in both of molecules A and B
Q^2	Predictive squared correlation coefficient
r	Correlation coefficient in the correlation matrix
R_s	Resolution
R^2	Coefficient of determination
s	Standard error of estimation
s_0	Slope of the regression line through the origin
t_0	Void time

List of included publications

Peer-reviewed articles

1. S.H. Park, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl. Enhanced methodology for porting ion chromatography retention data. *J. Chromatogr. A* 1436 (2016) 59-63. (Chapter 3)
2. S.H. Park, P.R. Haddad, M. Talebi, E. Tyteca, R.I.J. Amos, R. Szucs, J.W. Dolan, C.A. Pohl. Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model. *J. Chromatogr. A* 1486 (2017) 68-75. (Chapter 4)
3. S.H. Park, M. Talebi, R.I.J. Amos, E. Tyteca, P.R. Haddad, R. Szucs, C.A. Pohl, J.W. Dolan. Towards a chromatographic similarity index to establish localised quantitative structure-retention relationships for retention prediction. II Use of Tanimoto similarity index in ion chromatography. *J. Chromatogr. A* 2017, in press (Chapter 5)
4. E. Tyteca, M. Talebi, R.I.J. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad. Towards a chromatographic similarity index to establish localised quantitative structure-retention models for retention prediction: use of retention factor ratio. *J. Chromatogr. A* 1486 (2017) 50-58. (Chapter 6)

Oral presentations at international conferences (presenting author underlined)

1. S.H. Park, P.R. Haddad, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, J.W. Dolan, C.A. Pohl. Retention prediction of inorganic anions and organic cations in ion chromatography based on quantitative structure-retention relationships, *International Symposium on Advances in Separation Science (ASASS 2016)*, 2016, Hobart, Australia.

2. S.H. Park, G. Schuster, M. Talebi, G.W. Dicinoski, R.A. Shellie, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl. Quantitative structure-retention relationships for inorganic ions in ion chromatography, *42nd International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2015)*, 2015, Geneva, Switzerland.
3. P.R. Haddad, S.H. Park, M. Taraji, Y. Wen, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, J.W. Dolan, C.A. Pohl. Role of structural similarity in prediction of retention in reversed-phase, ion-exchange and hydrophilic interaction liquid chromatography modes using quantitative structure-retention relationships, *44th International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2016)*, 2016, San Francisco, USA.
4. P.R. Haddad, S.H. Park, M. Taraji, Y. Wen, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, J.W. Dolan, C.A. Pohl. How important is structural similarity in quantitative structure-retention relationships, *40th International Symposium on Capillary Chromatography and 13th GCxGC Symposium*, 2016, Riva del Garda, Italy.
5. P.R. Haddad, S.H. Park, M. Taraji, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan. Use of quantitative structure-retention relationships to choose the optimal chromatographic techniques, *Pacifichem*, 2015, Honolulu, USA.
6. P.R. Haddad, S.H. Park, M. Taraji, Y. Wen, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan. Prediction of retention times in reversed-phase, HILIC and ion chromatography based on chemical structures of analytes, *43rd International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2015)*, 2015, Beijing, China.
7. R. Szucs, M. H-Brown, P.R. Haddad, S.H. Park, M. Taraji, G. Schuster, M. Talebi. The role of quantitative structure retention relationships in quality by design chromatographic method development: optimization of retention models, *42nd*

International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2015), 2015, Geneva, Switzerland.

8. P.R. Haddad, S.H. Park, M. Taraji, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan. Prediction of retention times in reversed-phase, HILIC and ion chromatography based on chemical structures of analytes, *14th Asia-Pacific International Symposium on Microscale Separations and Analysis (APCE 2014)*, 2014, Kyoto, Japan.
9. P.R. Haddad, S.H. Park, M. Taraji, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, R. Szucs, C.A. Pohl, J.W. Dolan. Prediction of retention behaviour in reversed-phase, HILIC and ion chromatography based on chemical structures of analytes, *30th International Symposium on Chromatography (ISC 2014)*, 2014, Salzburg, Austria.

Poster presentations at international conferences (presenting author underlined)

1. S.H. Park, P.R. Haddad, M. Talebi, R.I.J. Amos, R.A. Shellie, R. Szucs, J.W. Dolan, C.A. Pohl. Retention prediction of inorganic anions and organic cations in ion chromatography based on quantitative structure-retention relationships, *44th International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2016)*, 2016, San Francisco, USA.
2. E. Tyteca, S.H. Park, R.A. Shellie, P.R. Haddad, G. Desmet. Possibilities of computer-assisted multi-segment gradient optimization in ion chromatography, *42nd International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2015)*, 2015, Geneva, Switzerland.
3. S.H. Park, R.A. Shellie, G.W. Dicinoski, M. Talebi, C. Johns, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl. Updating methodology for "porting" ion chromatography retention

data, *Asia Pacific-Korea Conference on science and technology (AKC)*, 2014, Sydney, Australia.

4. S.H. Park, R.A. Shellie, G.W. Dicinoski, M. Talebi, C. Johns, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl. Retention prediction of inorganic anions in ion-exchange chromatography based on quantitative structure-retention relationships, *40th International Symposium on High Performance Liquid Phase Separations and Related Techniques (HPLC 2013)*, 2013, Hobart, Australia.

Abstract

The streamlined "scoping" of analytical methods, leading to reductions in time and cost, is one of major challenges in chromatographic method development in the pharmaceutical and other industries. The screening (or scoping) of initial chromatographic parameters, such as stationary phase candidates and rough mobile phase conditions, can be accelerated by the *in-silico* prediction of chromatographic retention. Quantitative Structure-Retention Relationships (QSRRs), which allow the prediction of retention time of analytes based only on their chemical structures, offers an attractive approach to speed up the scoping phase of method development by minimising the number of initial experiments required at this stage. This present study presents an investigation on the establishment of QSRR models, leading to rapid, robust, and rugged scoping method development in ion chromatography (IC), which is a well-established technique for the analysis of inorganic and organic ions.

This study commenced with updating, by a new "porting" methodology, the retention data of anions embedded in the Virtual Column® Software (Thermo Fisher Scientific, Sunnyvale, CA, USA), which is a commercial tool for simulating and optimising IC separations. The retention data of inorganic and small organic anions on three Thermo Fisher Scientific columns (49 ions on AS20, 41 on AS19, and 40 on AS11HC) were recalibrated by the porting equations, which express the relationship between old (or embedded) and new (or updated) retention data. The three porting equations on each column under study were derived by performing three isocratic separations with six representative analyte ions (chloride, bromide, iodide, perchlorate, sulfate, and thiosulfate). Average errors in retention times for ten test anions under three eluent concentrations on the three columns were found to be less than 1.3%, showing that the updated retention data were accurate and reliable enough to be employed as anion datasets for the subsequent QSRR studies.

The QSRR approach was integrated with a well-known linear solvent strength (LSS) model in IC, which correlates retention factor with eluent concentration: log$k$ = a – $b$log[eluent]. The prediction of the two retention parameters (a and b) in this model has a great advantage in that it would allow the prediction of retention times (t_R) of ions under all eluent compositions. The first QSRR study was performed on the abovementioned updated retention data of inorganic and small organic anions. Molecular descriptors were calculated for each test analyte and an evolutionary algorithm (EA) was employed to extract the most relevant molecular descriptors, followed by multiple linear regression (MLR) to generate QSRR models correlating the two parameters (a and b) with the selected molecular descriptors. Six QSRR models (a and b, respectively, on the three columns AS20, AS19, and AS11HC) were successfully generated, resulting in good predictive performance for a- and b-values (R^2 >0.98 and root-mean-square error (*RMSE*) <0.11 for training sets; $Q_{ext(F3)}^2$ >0.7 and root-mean-square error of prediction (*RMSEP*) <0.4 for external test sets), and hence t_R-values (R^2 of 0.98, *RMSE* of 0.89 min, $Q_{ext(F3)}^2$ of 0.96 and *RMSEP* of 1.18 min).

The main objective in QSRR modelling is to build models with high predictive power, allowing reliable retention prediction for the unknown compounds across the chromatographic space. With the aim of enhancing the predictive power of the models, an approach called "federation of local models" was employed in generating QSRR models, where, for each target ion, a local model is created using only either structurally or chromatographically similar ions from the dataset, as opposed to the classical approach where all the compounds in the dataset are used to derive a “global” predictive model. The role of similarity in QSRR in IC was investigated systematically by employing a dataset (on Thermo Fisher Scientific CS17 column), consisting of larger molecular weight cations (molecular mass up to 507), being of interest in the pharmaceutical industry, along with the anion datasets used above. The QSRR models for a- and b-values were developed by employing a genetic algorithm-partial least

squares (GA-PLS) regression tool for descriptor selection and modelling. This approach was preferred to MLR in order to minimise problems of chance correlation, over-fitting, and multi-collinearity of descriptors associated with modelling datasets with a much greater number of descriptors than samples. For this purpose, similar ions to each target ion were clustered into training sets prior to QSRR modelling, by utilising various similarity measures such as Tanimoto similarity (TS) index, the retention factor ratio (or *k*-ratio), or the combination of TS and *k*-ratio.

First, the pair-wise TS score was utilised. The TS is a popular similarity measure based on a comparison of 2-dimentional fingerprints existing between two molecules and the TS score varies between 0 to 1, with 1 representing 100% similarity. When applying a TS threshold of 0.6, predictive and reliable QSRR models were produced ($Q_{ext(F2)}^2$ > 0.8 and *RMSEP* < 0.1), and hence accurate retention time predictions with a *RMSEP* of 0.44 min. Second, the *k*-ratio, proposed as a chromatographic similarity index, was investigated. Ions eluted in the vicinity of any given target ion would be assumed to be chromatographically similar to the target ion and these ions would exhibit *k*-ratio values close to 1. In addition to successful QSRR modelling of inorganic and small organic anions on two columns (AS20 and AS19), the *k*-ratio-based ion clustering provided the most accurate QSRR models for cations on CS17 column, with excellent predictive power for retention time prediction ($Q_{ext(F2)}^2$ of 0.99 and *RMSEP* of 0.22 min). However, the *k*-ratio method is impractical for retention prediction of unknown ions due to the lack of the retention information for those ions. In order to address this limitation, a dual filter-based localised QSRR modelling approach was developed, where a *k*-ratio (<1.2) index was combined with both a TS (>0.5) and a $\Delta \log P$ (<0.4) index. Predictive QSRR models for *a*-, *b*- and t_R- values were created successfully ($Q_{ext(F2)}^2$ of 0.96, 0.95, and 0.96, *RMSEP* of 0.06, 0.02, and 0.38 min). Additionally, the dual filter-based clustering method allowed more ions (50 ions) to be eligible for QSRR modelling, in comparison to the TS approach (22 ions).

Table of contents

Chapter 1

Literature review

1.1. Ion Chromatography

Ion chromatography (IC) is a well-established analytical technique for the determination of ionic species, predominantly using ion-exchange as the separation method [1]. IC has been typically used for the separation of inorganic and low molecular weight organic ions and also extensively employed for the analysis of a wide range of ionic and ionogenic species, including larger molecular weight organic ions, carbohydrates, amino acids, peptides, and proteins [2].

Retention in IC is governed by the interactions of analyte ions, both with the ion-exchange groups on the stationary phase and with the mobile phase (or eluent) [1]:

— **(1.1)**

where k is the retention factor, w_s is the weight of the stationary phase, V_m is the volume of the eluent, and D_A is the distribution of the analyte ion between the stationary phase and the mobile phase. The D_A term is defined as a ratio of the concentrations of analyte ions (denoted as A) in the stationary phase to that in the mobile phase and is given by [1]:

— **(1.2)**

Analytes having higher D_A possess higher affinity to the stationary phase than the mobile phase **[Eq. (1.2)]**, and hence larger k values **[Eq. (1.1)]**, implying longer retention times of those analytes. From **Eq. (1.1)** and **(1.2)**, stationary phase and mobile phase compositions are critical parameters affecting the retention of analytes in IC and hence, these chromatographic conditions need to be correctly selected for a

desired separation (This will be discussed in detail in section 1.2). In addition, a detection method, generally depending on properties of the analyte and the eluent, is also an important parameter to be determined [1, 3-5]. The discussion of detection methods in section 1.1 will be limited to two main detection modes employed through the present study: suppressed conductivity detection and UV detection.

1.1.1. Stationary phases

Ion-exchange stationary phases are largely subdivided into two types: silica-based and polymer-based. Polymer-based stationary phases, used throughout this work, are predominantly employed in IC since organic polymers (used as the support material of the ion-exchange stationary phase) are stable in extreme pH, facilitating the use of a hydroxide eluent (the typical eluent for anionic analysis) even at a high concentration [6]. These polymer-based stationary phases are mostly compatible with organic solvents, which has facilitated the determination of organic ionogenic analytes using IC [7]. More solvent-tolerant ion-exchange columns can be produced by increasing the degree of cross-linking of substrate beads to which the ion-exchange functional groups are attached and thus, decreasing the polymer swelling due to the addition of organic solvent. A stationary phase with a degree of cross-linking greater than 50% has allowed the use of all common HPLC solvents [7].

Ion-exchange stationary phases largely consist of five main substrate structures [8]: electrostatic agglomerated ultrawide-pore substrates (type 1), polymer-encapsulated substrates (type 2), chemically derivatised polymeric substrates (type 3), polymer-grafted film on porous substrates (type 4), and step-growth polymers on polymeric substrates (type 5). Latex-agglomerated ion-exchangers (corresponding to type 1, above) are one of the most common polymeric ion-exchange stationary phases, and a wide range of latex agglomerated resins (including AS11HC column used in this work, as well as AS4, AS9HC, AS14 and AS15) have been developed by Thermo Fisher

Scientific (previously Dionex) [3]. Typically, this type of stationary phase is manufactured by the following procedure: resins based on microporous (gel-type) styrene-divinylbenzene copolymers are functionalised with sulfonate groups and then aminated latex particles (porous polymer beads with cationic ion-exchange functional groups) are agglomerated on this surface [1, 4]. Generally, quaternary ammonium groups are used as anion-exchange functional groups; and sulfonate, phosphonate, or carboxyl groups are used for cation-exchange [1, 3]. Due to the small particle size of the latex and the thin coating layer, highly efficient separations can be obtained using this type of column [4]. In addition to polystyrene-divinylbenzene copolymers, polymethacrylate and polyvinyl resins are also widely used as substrate materials for polymer-based stationary phases [3, 6].

The cation columns developed by Thermo Fisher Scientific (such as CS12, CS18 and CS19) fall into the type 4 category which is comprised of high capacity packing materials [8, 9]. This type of resin is made by attaching polymer strands on the surface of a substrate [9]. The substrate is prepared either by forming polymerisable groups on the surface or by modifying the surface enabling the introduction of polymerisable groups. Subsequently, resin, monomer, and initiator are reacted to generate the grafted composite. Since the cross-linking monomers are incorporated in the reaction mixture to form a gel, with substrate particles suspended in the reaction mixture, the cross-linker in this type of stationary phase cannot be used for selectivity control (unless it is added after the graft step). The CS19 column is optimised for use with inorganic cations and aliphatic amines, and the CS18 for the separation of common inorganic cations, polar amines including biogenic amines, and multiply charged species. These columns do not require organic solvent in the eluent for the separations of the analytes described above.

Newly developed columns such as AS22, AS23, and AS25, along with the more than 10 anion-exchange columns manufactured continuously over the past decade (such as

AS19 and AS20), are type 5 stationary phases [8]. This type of resin is produced by the combination of synthesis approaches for type 1 and type 4 stationary phases (*i.e.* the formation of a wide-pore polymeric surface followed by the electrostatic attachment of a hyper-branched condensation polymer to the substrate surface). The wide-pore (surface-sulfonated) substrate is functionalised with anion-exchange groups in the same manner as the latex-agglomerated ion-exchangers (*i.e.*, the production of an amine-rich "basement" polymer bound electrostatically to the resin surface) and an epoxy-amine copolymer is then formed in the presence of the "basement" polymer, by alternating the diepoxy monomer and the primary amine. Branch sites can be introduced by adding a second primary amine, following the addition of the second diepoxy monomer. Through the alternating treatments (of diepoxy monomer and primary amine), a novel hyper-branched polymer grows off the substrate surface. The column capacity increases as the number of alternating reaction cycles increase. In addition, these hydrophilic columns are completely compatible with hydroxide eluents even at high pH, due to the presence of the aliphatic substituents [10].

There are important parameters imparting different properties to the stationary phase, which include the ion-exchange group structure, resin capacity, and the diameter and the particle size of the column. The ion-exchange functional group structure (*e.g.*, variation in quaternary ammonium group on the latex) can influence the selectivity [4]. When the methyl group in the quaternary ammonium group was replaced with ethanol groups (such as methyldiethanolamine [MDEA], dimethylethanolamine [DMEA]) the eluting power of the hydroxide eluent was found to be considerably enhanced [11]. Additionally, as the size of the alkyl groups on the quaternary nitrogen increased, polarisable anions (such as nitrate, chlorate, and iodide ions) were eluted later [4]. Similarly, the resolution of five anions (formate, chloride, bromide, nitrate, and chlorate) was improved upon an increase of the size of alkyl

groups (*e.g.*, from trimethylamine [TMA] to tripropylamine [TPA] and trihexylamine [THA]) [12].

Low-capacity stationary phases, consisting of nonpolar macroporous copolymers with relatively few ion-exchange sites provide short analysis times. When using a low-capacity column, an eluent of lower concentration is generally employed [4].

The diameter of the column (or stationary phase) is also a significant factor in the separation of analytes. Recently, columns with smaller inner diameters (such as microbore [1-2 mm i.d.] and capillary [<1 mm i.d.] columns) have been increasingly used, due to two main advantages: higher sensitivity (or higher peak height) with an injection of the same mass of sample, and shorter analysis time at the same flow-rate [4, 8]. In addition, smaller concentrations of samples can be analysed and the flow-rate can be decreased while still giving the same elution time, resulting in lower solvent content [8].

The particle size of columns influences the separation efficiency. Columns having smaller particles provide efficient separations, facilitating the use of shorter columns, leading to faster analysis [8]. Thus, commercial columns based on reduced particle sizes have been increasingly developed for fast analysis, including the TSKgel SuperIC-Anion HS column (Tosoh Bioscience, 100 mm x 4.6 mm, 3.5 μm), the IC SI-35 4D column (Shodex, 150 mm x 4 mm, 3.5 μm), and the Thermo Fisher Scientific IonPac AS18 (4 μm), AS11-HC (4 μm), and CS19 (4 μm).

1.1.2. Eluent

The introduction of the electrolytic eluent generator to the IC system has led to the generation of hydroxide eluents with accurate and precise concentrations, without carbonate contamination, and with reproducible gradient elution [2]. Typically, dilute bases such as hydroxide or carbonate/bicarbonate are used as the eluent for anion analysis, and acids such as methanesulfonic acid (MSA) for cation analysis [3]. The

eluent (E) takes part in the separation of analytes by competing with the analyte ions (A) for the same ion-exchange sites on the stationary phase and the reaction for two monovalent anions (as an example) is [1, 13]:

$$A_m^- + E_s^- \rightleftharpoons A_s^- + E_m^- \quad \textbf{(1.3)}$$

where the subscripts m and s denote the mobile phase and the stationary phase, respectively. The equilibrium constant of the reaction is therefore given by:

$$K_{A,E} = \frac{[A^-]_s[E^-]_m}{[A^-]_m[E^-]_s} \quad \textbf{(1.4)}$$

where the $[A^-]$ and the $[E^-]$ indicate the concentrations of the analyte ions [in mmol/l] and the competing eluent ions [in mmol/l], respectively. The equilibrium constant $K_{A,E}$ is referred to as the selectivity coefficient and retention increases upon an increase of $K_{A,E}$ [1]. The selectivity coefficient of the eluent and the analyte ion can be useful when selecting the eluent [3]. Eluent ions (competing ions) having high $K_{A,E}$ values are often preferred as the eluent, due to their excellent elution power even in dilute solution. In the case of analyte ions that are eluted too early, retention can be controlled either by lowering the ionic strength of the eluent or changing the eluent (to one having a smaller selectivity coefficient).

The eluent composition is a critical parameter influencing the retention of the analytes. Sub-parameters determining the eluent composition include eluent concentration, eluent type, and the organic modifier content [1, 4]. The retention times of analytes decrease as the eluent strength increases. In other words, the use of a higher eluent concentration and an eluent ion with a higher charge (such as carbonate) can result in a shorter analysis time [1, 4]. Additionally, it may be useful to add an appropriate amount (*e.g.*, 10 % to 20 %) of an organic modifier (such as methanol and acetonitrile) to the eluent, to completely dissolve sample components and to improve the compatibility of analyte samples with the ion-exchange stationary phase, as well as to facilitate column clean-up [7, 14]. Worth noting is that the peak tailing for

polarisable anions may also be lessened by the use of the organic modifier in the eluent [4].

Numerous studies using the eluent modified by the addition of organic solvent have been reported for the determination for organic ionogenic analytes which are separated based on mixed-mode retention including both electrostatic and hydrophobic interactions [7, 15-17]. Zakaria *et al.* [16] has demonstrated the effect of methanol on the separation of anionic analytes of pharmaceutical interest. Poor resolution was observed where no organic modifier was added, whereas better resolution and peak shape, as well as a reduction of retention were obtained as the methanol content increased. Similarly, another study for the analysis of tetracyclines (such as tetracycline, chlortetracycline, doxycycline, and oxytetracycline) has shown a decrease in the retention time upon an increase of acetonitrile content [18].

To introduce the organic modifier to the typical IC system, the system configuration needs to be modified since the eluent generator (or more specifically, the membrane in the Elu-Gen cartridge) in the IC system cannot tolerate organic solvents [16]. Briefly, the hydroxide eluent can be generated from the eluent generator (EG) and then mixed with methanol at a T-piece. For an IC-mass spectrometry (MS) system, organic modifiers can be introduced either by the configuration described above (*i.e.*, by pre-column addition), or after the suppressor (*i.e.*, by post-suppressor addition) where the presence of organic solvents improves the volatility of the eluent, which in turn increases the detection sensitivity due to efficient desolvation/ionisation processes in the electrospray ionisation (ESI) interface [19].

Finally, the type of elution mode (isocratic, gradient, or multi-step gradient elution) is also an important parameter influencing the separation of analytes. Isocratic elution is the simplest elution mode where a constant eluent composition is used throughout the entire separation. However, there are some disadvantages in the use of this type of elution. First, analytes with small and medium distribution coefficients can be co-eluted

or show poor resolution at high eluent concentrations. Second, peak capacity can be smaller compared to that given with gradient elution. Finally, the separation of the later eluting analytes can require a long analysis time [1, 20]. Gradient elution - varying the eluent composition during the chromatographic run - is a powerful tool to overcome the limitations described above. Additionally, when analyte samples are very complex or the difference in retention times between early- and later-eluted ions is sizeable, implementing gradient elution is desirable [4, 21]. There are two modes for gradient elution: linear gradient elution and multi-step gradient elution. In the former mode, the eluent concentration is varied linearly over time, and the latter mode generally comprises a cascading run of isocratic and linear gradient elution.

1.1.3. Suppressed conductivity detection

Since ionic species are electrically conducting, conductivity detection (CD) is frequently selected as the detection method. CD is based on the measurement of the conductance, generated by a voltage applied across two electrodes in the detector cell. The background conductance G of the eluent, changes when an analyte ion passes the detector. The change in conductance ΔG due to the analyte ion provides an analyte peak in the chromatogram. For anion-exchange, ΔG (in μS) is given by [1, 22]:

$$\Delta G = \left[\left(\frac{(\lambda_{E^+} + \lambda_{A^-})\cdot\alpha_A - (\lambda_{E^+} + \lambda_{E^-})\cdot\alpha_E\cdot\alpha_A}{10^{-3}\cdot K}\right)\right]\cdot C_A \qquad \textbf{(1.5)}$$

where λ_{E^+}, λ_{E^-}, and λ_{A^-} are the equivalent ionic conductances of the eluent cations, eluent anions, and analyte anions, respectively, α_A and α_E are the degree of dissociation of the analyte and the eluent, C_A is the analyte concentration, and K is the detector cell constant. According to **(Eq. 1.5)**, the detector signal is proportional to the analyte concentration. In addition, the peak heights (or areas) of analytes with the same concentration depend on the equivalent ionic conductance of the eluent and the analyte ions, as well as the degree of dissociation of eluent and analyte ions. The equivalent ion conductance measurements for common ions are available from the

literature [1] and the degrees of dissociation of eluent and analyte ions are determined by the eluent conditions (such as pH) [3] and analyte p*K*a.

In most cases, higher concentrations of eluents (10^4 - 10^5 higher than the analyte) are used for the separation of analytes in IC, which results in low detection sensitivity. Since the introduction of a suppressor to the IC system [23], this limitation has been addressed, *i.e.*, the detection sensitivity of analytes has been significantly enhanced by decreasing the background conductivity by the eluent suppression.

Anion suppression of a KOH eluent, as an example, is performed based on the following two cation-exchange reactions, taking place in the anion suppressor which is located between the separation column and the conductivity detector [22]:

$$\text{Resin} - \text{H}^+ + \text{K}^+\text{OH}^- \rightleftharpoons \text{Resin} - \text{K}^+ + \text{H}_2\text{O} \qquad \textbf{(1.6)}$$

$$\text{Resin} - \text{H}^+ + \text{K}^+\text{A}^- \rightleftharpoons \text{Resin} - \text{K}^+ + \text{H}^+\text{A}^- \qquad \textbf{(1.7)}$$

According to **Eq. 1.6**, the counter-ion in the eluent (K^+) is replaced with the hydronium ion (H^+). This hydronium ion produces water with the hydroxide ion so that the background conductivity of the eluent, which corresponds to the noise, is decreased. Similarly, the counter-ion in the analyte band (K^+) is replaced with the hydronium ion **(Eq. 1.7)**. This hydronium ion is more conductive than the corresponding potassium ion so that the response (or conductivity signal) for the analyte is increased. As a result, detection sensitivity is enhanced by increasing the ratio of signal-to-noise. The cation suppression takes place by a similar procedure to the anion-exchange reactions.

1.1.4. UV detection

Ultra-violet (UV) detection can be another common choice for IC separations when analytes have UV-absorbing chromophores. There are two types of detection modes: direct and indirect UV detection. Direct UV detection which was used in this work analyses only compounds having chromophores that absorb UV light, where the absorbance of the analyte and the eluent (A_{sample} and $A_{background}$) follow Beer's law and

the UV absorbance change on elution of a analyte (*i.e.*, ΔA) is measured by the detector (as an example, for anion-exchange under conditions where the (singly charged) analyte is fully ionised [or dissociated]) as follows [1, 5]:

$$\Delta A = A_{sample} - A_{background} = (\varepsilon_{S^-} - \varepsilon_{E^-})C_s m \qquad \textbf{(1.8)}$$

where ε_{S^-} and ε_{E^-} are the molar absorptivity of the sample (*i.e.*, analyte) and the eluent anion, C_s is the analyte concentration, and m is the path-length of the detector cell. As seen in **Eq. 1.8**, the detector response increases upon increasing the analyte concentration and numerous studies analysing organic ions, where their molar absorptivities are larger than that of the eluent, have employed the direct UV detection [5, 24]. On the other hand, UV-transparent ionic species can be determined by indirect UV detection provided that an eluent containing UV-absorbing counter-ions is employed (*i.e.*, molar absorptivity: eluent > analytes) [5, 24].

Numerous studies using direct UV detection have been reported for the analysis of pharmaceutically-related organic ionogenic compounds [14, 17, 25-27]. In IC, direct UV detection may be employed in either suppressed or non-suppressed modes. For example, organic anions and cations with larger molecular weights (ranging from 95 to 384) [14, 16] were analysed on commercial columns (Thermo Fisher Scientific AS24, AS20, AS16, AS11HC and AS11) using ion suppression, at 190 nm, 220 nm and 254 nm depending on the characteristics of analytes. Although several examples have successfully employed suppressed UV detection for the separations of pharmaceutical compounds, by locating the UV detector after the suppressor in a typical IC system [14, 16], non-suppressed UV detection is probably the most straightforward detection method for the chromophoric compounds [18, 28-30]. Various classes of pharmaceutical compounds, including beta-blocker drugs (270 nm) [28], paracetamol and salicylic acid (300 nm) [30], tetracycline antibiotics (350 nm) [18], and flucloxacillin and amoxicillin (225 nm) [29] have been determined by direct UV

detection. In addition, the instrument for UV detection can be either a general HPLC instrument or a typical IC instrument since the suppressor is not required.

Finally, UV-transparent ionic compounds may also be analysed using direct UV detection after the use of appropriate derivatisation techniques. For example, polyvalent anions (such as polyphosphate, polyphosphonic and polyphosphinic acids) could be detected at 330 nm by derivatising these anions with iron (III) nitrate [31].

1.2. Analytical method development

Method development in liquid chromatography is defined as the process of determining the correct chromatographic parameters leading to successful analysis of the target compounds. Chromatographic method development typically consists of two main steps: the scoping of methods, followed by the optimisation of the chosen method [32]. The scoping (or screening) step is the process of identifying appropriate starting conditions for experiments (*i.e.*, the selection of rough chromatographic parameters such as the column type and broad mobile compositions), and the optimisation step is the process of determining the precise method parameters for establishing the final method [33]. The traditional trial-and-error approach for method development has been increasingly replaced with computer-assisted method development which generally employs retention models, simplifying the selection of method parameters. Computer-assisted method development accordingly reduces the time and cost required and increases the robustness and ruggedness of the final method [32, 34]. Numerous computer software packages have been commercialised for both the scoping and the optimisation processes of method development. ACD software (Advanced Chemistry Development, Toronto, Canada) and ChromSword (Merck KGaA, Darmstadt, Germany) are typical examples of the software packages for scoping [34], and Virtual Column (Thermo Fisher Scientific, Sunnyvale, USA) [35] and DryLab (LC Resources, USA) [36] are used for the optimisation of chromatographic methods. These packages

can determine both the starting and the optimal chromatographic conditions by constructing and comparing all possible separations within a chosen experimental space of columns and eluents. Additionally, this procedure is claimed to be simple, robust and easily applicable for various systems [37]. This section highlights the introduction of three representative IC retention models embedded in the Virtual Column software (section 1.2.1), rapid "scoping" method development using commercial software packages (section 1.2.2), which is an ultimate future goal of this project, and the computer-assisted optimisation procedure based on the Virtual Column software, from which embedded anion data were used in this work (section 1.2.3).

1.2.1. Retention time modelling

The generation of retention models for desired analytes facilitates the search for either starting or optimal conditions in chromatographic method development by calculating (or predicting) retention times and then comparing these with experimental retention times of the analytes to be separated. Among the numerous retention models developed in IC, several important isocratic and gradient eluent retention models are discussed in this section. Model equations for the retention of anions are introduced for simplicity. The models for cations, while not outlined here, are similarly derived.

There are two major models - the Linear Solvent Strength (LSS) model and the LSS model-empirical approach (LSSM-EA) - that are used for retention prediction under isocratic eluent conditions. The LSS model is a fundamentally-derived isocratic retention model in IC, providing a quantitative relationship between the retention factor and some measurable stationary phase and eluent parameters [1]. For a single-component eluent (with one competing ion such as hydroxide), the LSS model is [1]:

$$\log k = \frac{1}{y}\log\left(K_{A,E}\right) + \frac{x}{y}\log\left(\frac{Q}{y}\right) + \log\left(\frac{w}{V_m}\right) - \frac{x}{y}\log[E^{y^-}] \qquad \textbf{(1.9)}$$

where k is the retention factor, $K_{A,E}$ is the ion-exchange selectivity coefficient between the analyte and the eluent competing ion, w and Q are, respectively, the mass and the effective ion-exchange capacity of the stationary phase, V_m is the dead volume of the stationary phase, x is the charge on the analyte, y is the charge on the eluent, and $[E^{y-}]$ is the concentration of the eluent.

For a given column, eluent, and solute ion, $K_{A,E}$, w, Q, and V_m are constants and **Eq. 1.9** can therefore be simplified to:

$$\log k = a - b\log[E^{y-}] \quad \textbf{(1.10)}$$

where a relates to the degree of interaction between analytes and the stationary phase, and b indicates the ratio of the effective charges of the analyte and the eluent competing ion. **Eq. 1.10**, showing the linear relationship between the logarithm of the retention factor and the logarithm of the eluent concentration, is often referred to as the LSS model in the IC literature [16, 17]. This model has been used throughout this study. Generally, the two retention parameters (a and b) are estimated by plotting $\log k$ against $\log[E^{y-}]$, by performing several (usually three) isocratic runs [17, 38-40].

The LSSM-EA is derived by extending the LSS model for a dual-component eluent (with two competing ions such as carbonate/bicarbonate), based on a 2-dimensional design of experiments consisting of mole fraction ($R = [E^{2-}]/[E_T]$) of the eluent species of higher selectivity coefficient (one dimension) and the total concentration of the two eluent components $[E_T]$ (the other dimension) [35, 38]:

$$\log k = (f_1 + f_2[E_T]) + (f_3 + f_4[E_T])\log[E^{2-}] \quad \textbf{(1.11)}$$

where $[E^{2-}]$ is the eluent concentration of the doubly charged ion. Four prediction coefficients (f_1, f_2, f_3, f_4) are determined experimentally. When the eluent concentration of doubly charged competing ions $[E^{2-}]$ is zero (*i.e.*, in presence of only singly charged competing ions), **Eq. 1.11** becomes **Eq. 1.10** [38].

The Rocklin model [41] and the Jandera model [42] are both widely investigated gradient retention models in IC. The following Rocklin model, along with the LSS model

and LSSM-EA, has been embedded in the Virtual Column Software, which is a commercial optimisation tool for IC separations and will be discussed in detail in section 1.3.3:

$$\log k = \log C_g - \frac{x}{x+y}\log R_g \quad \textbf{(1.12)}$$

where R_g is the gradient ramp (mM/min), y the charge of the eluent, x the charge of the analyte and C_g is a gradient constant. The C_g and slope [x/(x+y)] are determined experimentally by plotting log$k$ against log R_g. The model is valid for single eluent species. Since the model does not include important gradient parameters, such as the initial eluent concentration and the eluent flow-rate, it is limited in that different gradient constants and slopes are required for different initial eluent concentrations in order to predict the gradient retention.

The Jandera model can be used for more general prediction of retention times of analytes under gradient conditions by including the initial eluent concentration in the model, and is given as follows:

$$t_g = \left(\frac{1}{F_v}\right)\left\{\left(\frac{1}{B}\right)\left[(zb_i + 1)Ba_i t_0 F_v + C_i^{(zb_i+1)}\right]^{1/(zb_i+1)} - \frac{C_i^{1/z}}{B}\right\} + t_0 \quad \textbf{(1.13)}$$

where t_g is the gradient retention time of the analyte, F_v the eluent flow-rate (mL/min), B the normalised gradient ramp ($B = R/F_v$ [mM/mL]), z the adjustable parameter related to the shape of gradient ramp (z =1 for a linear gradient), C_i the initial concentration for the gradient, t_0 void time, and a_i and b_i represent the same values as a and b in **Eq. 1.10**. Shellie *et al.* [40] have evaluated this model by two approaches: gradient data for gradient prediction (GDGP) and isocratic data for gradient prediction (IDGP). The IDGP approach used a_i and b_i, estimated by the LSS model based on three isocratic runs, whereas for the GDGP approach, a_i and b_i were calculated by **Eq. 1.13** using six combinations of B and C_i (two gradient ramp and low, intermediate, and high initial gradient eluent concentration) for a fixed analyte, column, and the eluent flow-rate u. Using these a_i and b_i values, the gradient retention times were predicted and

then compared with the corresponding observed retention times. The prediction errors by the two approaches were comparable [40]. It was further concluded that a_i and b_i values determined using the LSS model can predict analyte retention times under a multi-step gradient eluent profile, as well as for isocratic and linear gradient eluent conditions [40].

1.2.2. Scoping of chromatographic methods

The correct identification of rough chromatographic parameters, including the best chromatographic mode (such as reversed-phase liquid chromatography [RPLC], hydrophilic interaction liquid chromatography [HILIC], or ion chromatography [IC]), stationary phase, and the broad mobile phase conditions, is the main goal for the scoping procedure in method development [32]. Among various parameters described above, the selection of an appropriate column having the desired separation selectivity (α) -leading to good separation and acceptable peak shapes of the analytes- is one of the biggest challenges during this step. For screening experiments, it is practically useful to select test columns, maximising differences in column properties and minimising the number of columns explored, to cover as much of experimental space as possible [33]. Van Gyseghem *et al.* [43] have evaluated 28 stationary phases to develop "scoping" methods for the separation of impurities in drug substances (*i.e.*, to define starting conditions for screening experiments). Various chemometric approaches (such as principal component analysis (PCA), colour maps, dendrograms *etc*.) were utilised, to select orthogonal chromatographic systems.

For rapid scoping, several studies employing commercial software packages such as Column Match™ (Rheodyne LLC, CA, USA) and ChromSword (Merck KGaA, Darmstadt) have been reported [33, 34, 44-46]. These software packages have been developed for RPLC method development and there is no available commercial software package for scoping method development in IC. The Column Match™ is a

useful tool, with embedded databases of numerous reversed-phase columns, whereby the orthogonality or equivalency of two columns can be determined by using a function of column selectivity F_s based on five parameters (hydrophobicity, sterics, hydrogen bonding acceptance and donation, and ionic charge) [33, 47]. Biswas *et al.* [33] have developed a simple approach for choosing scout columns, based on hierarchical clustering using data available in Column Match™. The clustering of 68 RPLC columns was performed on the above five parameters, and a subset of columns, showing maximally different properties, was then selected as the scout column set for successful method development for the RPLC separation of pharmaceutical compounds and their related impurities. Krisko *et al.* [45] has used the Column Match™ software to select 16 different stationary phases varying in hydrophobicity and polarity, to develop a RPLC separation method for drug candidates (including mixtures of bupivacaine and its metabolites, atenolol, nitrendipine, and their degradation products). Data for the resulting 16 columns, along with another data (such as mobile phase pH and test analytes) were then imported into DryLab™ software (LC Resources, CA, USA) to find optimal chromatographic conditions.

Xiao *et al.* [44] have employed ChromSword to screen mobile phase conditions and optimise the RPLC separation of structural epimers (betamethylepoxide and alphamethylepoxide). For the scoping of the most promising column, the authors used an LC Spiderling automated column switching system (Chiralizer Services, Newtown, PA, USA), by which, despite similar selectivity values ($1.05 < \alpha < 1.09$) for most of the columns under study, the YMC Hydrosphere C18 column was selected for the further optimisation of methods. A careful investigation of the total number of peaks, peak shapes, and the separation of small peaks from the main peaks of target analytes was used to choose the RPLC column with the highest resolution (or separation) power. The selection of starting conditions using the ChromSword software is based on a theoretical retention model, $\ln k' = a(V)^{2/3} + b(\Delta G) + c$. The software calculates

automatically the parameters V (partial molar volume in water) and ΔG (energy of interaction with water), based on the analytes' structures and embeds a database of the parameters (a, b, and c) on a wide range of commonly used RP columns, calculated using the sorbent/eluent system with an appropriate set of reference standards [48, 49].

1.2.3. Optimisation of chromatographic methods

Precise chromatographic parameters (eluent composition, pH, and flow-rate, temperature, stationary phase, *etc.*) are determined in the optimisation step. Systematic and computer-assisted optimisation procedures generally consist of the following main steps: (i) definition of the optimisation search area (minimum and maximum boundaries for parameters to be optimised, such as eluent concentration, eluent pH, initial gradient eluent concentrations, gradient ramp, organic solvent content); (ii) construction of experimental designs for acquiring the retention data (*e.g.*, full and fractional factorial designs, (D-)optimal designs and central composite designs *etc.*); (iii) collection of the retention data; (iv) evaluation of potential separations within the search area using suitable criteria (such as resolution and analysis time); (v) construction of a response surface; and (vi) identification of the optimal conditions using an optimisation criterion [1].

The resolution, used as an optimisation criterion, is [3, 38]:

$$R_S = 2\frac{t_{R_1} - t_{R_2}}{w_1 + w_2} \qquad \textbf{(1.14)}$$

where t_{R1} and t_{R2}, and w_1 and w_2 are the retention times of two adjacent peaks, and base widths of those peaks, respectively. It is worth noting that w_1 and w_2 are often replaced by the peak widths at half height and the resolution value is changed accordingly. As shown in **Eq. 1.14**, the construction of accurate models for both retention times and peak widths is important for accurate resolution prediction, leading to the precise optimisation of desired parameters. The resolution of a critical peak pair (or the least-

resolved pair of analytes) in a chromatogram has frequently been used as a useful criterion to evaluate the potential separations, based on a threshold value (*e.g.*, $R_s \geq 1.5$). The chromatograms satisfying the criterion are ranked by an additional criterion such as analysis time, to identify the optimal conditions [38].

On the basis of the steps described above, a commercial optimisation software package for IC separations - Virtual Column (Thermo Fisher Scientific, Sunnyvale, CA, USA) has been developed and is used routinely in the laboratory [35]. This software enables the simulation of possible separations within the search area, defined by the embedded retention databases, and then optimises the eluent compositions for the desired analysis. For isocratic separations using single-component eluents, the search area is defined as one dimension, consisting of the boundaries of the upper and lower limits of eluent concentrations. Accordingly, the response surface is shown as a plot of the optimisation criterion (*e.g.*, R_s) *vs.* eluent concentration. When using dual-component eluents, the search area is 2-dimensional, defined by two variables (*i.e.*, the total eluent concentration and the eluent molar ratio) and the response surface is therefore displayed as a contour plot (or a 3-dimensional surface) of the optimisation criterion *vs*. the two experimental variables (*i.e.*, the optimisation criterion on *z*-axis, and the total eluent concentration and the eluent molar ratio on the *x*- and *y*-axis). The following two criterion functions: the minimum resolution ($R_{s,min}$) and the normalised resolution product (r) are available for optimisation [35, 38]:

$$R_{s_{min}} = \min_{1 \to n-1} R_s \qquad \textbf{(1.15)}$$

$$r = \prod_{i=1}^{n-1} \frac{R_{s_{i,j+1}}}{\frac{1}{n-1}\sum_{i=1}^{n-1} R_{s_{i,j+1}}} \qquad \textbf{(1.16)}$$

where the subscripts i and j denote a specified peak and its adjacent peak, respectively. The $R_{s,min}$ indicates the resolution of the least-resolved pair of peaks (*i.e.*, the critical peak pair) and r becomes zero in the presence of co-eluted analyte ions. The suitable chromatograms for separation are determined based on either a desirable threshold

value of $R_{s,min}$ (*e.g.*, $R_{s,min} \geq 1.5$) or the maximum value of *r*. Finally, the best compromise of resolution and analysis time is then selected as an optimal condition.

1.3. Quantitative structure-retention relationship modelling

Quantitative structure-retention relationship (QSRR) modelling is a popular technique where chromatographic retention can be predicted from the molecular structures of analytes [50]. In QSRR modelling, chromatographic retention parameters (dependent [or **Y**-] variables) are generally expressed as a function of molecular descriptors (independent [or **X**-] variables) which encode the molecular structures of the analytes [50, 51]. QSRRs, which are statistically-derived mathematical relationships, have been applied to six major topics [52]: prediction of retention for a new analyte (which is the main goal of this study), identification of unknown analytes, selection of the most informative molecular descriptors (or feature selection), elucidation of molecular separation mechanisms (or retention mechanisms), characterisation and classification of chromatographic columns, and evaluation of complex physico-chemical properties of analytes (such as lipophilicity and dissociation constants) [52]. For the development of successful QSRRs, reliable input data (*i.e.*, chromatographic retention data for known chemical compounds having known molecular descriptors) and a rigid statistical analysis are essential requirements. Thus, this section highlights the introduction of essential elements in QSRR modelling: the definition and types of molecular descriptors (section 1.3.1), feature selection techniques to extract the most relevant features (or molecular descriptors) to the chromatographic retention (section 1.3.2), data collection and durability (section 1.3.3), and statistical modelling techniques with a focus on multiple linear regression (MLR) and partial least squares (PLS) regression (section 1.3.4). Finally, the role of similarity in QSRRs will be discussed in section 1.3.5.

1.3.1. Molecular descriptors

A molecular descriptor is defined as "*the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number (theoretical descriptor) or the result of some standardised experiment (experimental descriptor)*" [53]. Experimental descriptors are generally related to the physicochemical properties of molecules, including the dipole moment, the hydrophilic factor Hy, the log of the octanol/water partition coefficient log*P*, molar refractivity, and polarisability. Theoretical molecular descriptors are largely divided into five classes based on the characteristics of the molecular representation: zero-dimensional (0D)-, one-dimensional (1D)-, two-dimensional (2D)-, three-dimensional (3D)-, and four-dimensional (4D)-descriptors [53, 54]. The 0D descriptors (*e.g.*, constitutional descriptors) are derived based on information related to the chemical formula (such as the molecular mass, the number and type of atoms, and the sum of atomic van der Waals volume) and are independent from molecular conformations and connectivity. The 1D descriptors (*e.g.*, count descriptors) are based on the molecular representation, consisting of a list of molecular substructures (or molecular fragments such as functional groups, bonds and rings). The 2D descriptors (*e.g.*, total path count) contain the topological information of molecules, related to the bonding of atoms in a molecule (such as the bonding type and the interaction of particular atoms). The 3D-and 4D-descriptors are derived from a geometrical representation of a molecule and a stereo-electronic (or lattice) representation, respectively [54]. Examples of molecular descriptors commonly used in QSRR studies, are presented in **Table 1.1** [50]. The descriptors related to the molecular size are often found to be significant in QSRR studies, as the size changes the dispersive interactions (or London interactions) between analytes and the components of a chromatographic system. The separations of closely congeneric analytes are well described using molecular bulkiness descriptors. For example, the members of a homologous series can

Table 1.1 Exemplary structural descriptors used in QSRR studies.

Molecular bulkiness-related descriptors	***Molecular polarity-related (electronic) descriptors***
Carbon number Molecular mass Refractivity Polarisability van der Waals volume and area Solvent-accessible volume and area Total energy Calculated partition coefficient (CLOG P)	Dipole moments Atomic and fragmental electron excess charges Orbital energies of HOMO and LUMO Partially charged areas Local dipoles Submolecular polarity parameters
Molecular geometry-related (shape) descriptors	***Molecular graph-derived (topological) descriptors***
Length-to-breadth ratio STERIMOL parameters Moments of inertia Shadow area parameters	Molecular connectivity indexes Kappa indices Information content indexes Topological electronic index
Physicochemical empirical and semiempirical parameters	***Combined molecular shape/polarity parameters***
Hammett constants Hansch constants Taft steric constants Hydrophobic fragmental parameters Solubility parameters Linear solvation energy relationship (LSER) parameters Partition coefficient (LOG P) Boiling temperatures pK_a values	Comparative molecular field analysis (CoMFA) parameters Comparative molecular surface (CoMSA) parameters

be differentiated by the "carbon number" descriptors. Since changes in the polarity of analytes influence size-related interactions as well as dipole-dipole and hydrogen bonding interactions, it is also important to find appropriate molecular polarity descriptors such as "dipole moments" descriptors. The physicochemical (empirical and semi-empirical) parameters can be informative for the understanding of chromatographic retention mechanisms, although these descriptors for the analytes of QSRR-interest are often deficient [55]. Theoretical descriptors can be calculated by using commercial software such as Dragon. **Table 1.2** lists the Dragon descriptors with their block description and dimensionality [56-58].

1.3.2. Feature selection

For a typical analyte, the Dragon software (version 6.0, Talete, Milano, Italy), used in this study, generates more than 4000 theoretical descriptors and many of these descriptors are redundant or are irrelevant to chromatographic retention. Feature selection (or variable selection) is a process where a subset of the most significant descriptors (related to the chromatographic retention) is extracted from the large pool of descriptors given by the modelling software [59]. In QSRR, feature selection is a critical step required to create good quality QSRR models having both good fit and predictive power [59, 60]. Numerous feature selection techniques have been extensively reported [54, 59-62], including classification and regression tree (CART) [54, 61], stochastic gradient boosting for tree-based models (Treeboost) [61], random forests (RF) [61], variable iterative space shrinkage approach (VISSA) [60], uninformative variable elimination-partial least squares (UVE-PLS) [60, 61], genetic algorithm-multiple linear regression (GA-MLR) [59], and genetic algorithm-partial least squares (GA-PLS) [60, 62].

Among various feature selection techniques described above, the genetic algorithm (GA) has been frequently used for feature selection in QSRR modelling

Table 1.2 Descriptor blocks by the Dragon software (version 6.0) with the number of descriptors and their dimensionality [56-58].

No.	Block description	No. Desc.	Class
1	Constitutional descriptors	43	0D
2	Ring descriptors	32	2D
3	Topological indices	75	2D
4	Walk and path counts	46	2D
5	Connectivity indices	37	2D
6	Information indices	48	2D
7	2D matrix-based descriptors	550	2D
8	2D autocorrelations	213	2D
9	Burden eigenvalues	96	2D
10	P-VSA-like descriptors	45	2D
11	ETA indices	23	2D
12	Edge adjacency indices	324	2D
13	Geometrical descriptors	38	3D
14	3D matrix-based descriptors	90	3D
15	3D autocorrelations	80	3D
16	RDF descriptors	210	3D
17	3D-MoRSE descriptors	224	3D
18	WHIM descriptors	114	3D
19	GETAWAY descriptors	273	3D
20	Randic molecular profiles	41	3D
21	Functional group counts	154	1D
22	Atom-centred fragments	115	1D
23	Atom-type-E-state indices	170	2D
24	CATS 2D	150	2D
25	2D Atom Pairs	1596	2D
26	3D Atom Pairs	36	3D
27	Charge descriptors	15	Others
28	Molecular properties	20	Others
29	Drug-like indices	27	2D

[62-64], since the GA can produce efficiently the best subsets (with less redundant variables consisting of a lower number of informative features or descriptors) to generate successful models [65]. The GA, developed by Holland [66], is a representative class of evolutionary algorithms (EAs) which typically find an optimal solution for a complex problem (such as the selection of the best descriptors) based on the 'survival of the fittest' principle [67, 68]. In the GA, superior genes (or descriptors) leading to better results (or model responses) remain for the next generation while the worst ones are removed. The population therefore evolves toward the best solution (*i.e.*, the best subset of descriptors) through cross-over (or some sort of recombination) and mutation (or some random changes). Briefly, feature selection procedure by the GA consists of three main steps: initiation of the population (step 1), evaluation of the responses (step 2), and reproduction (step 3). The original population, comprised of a certain number of chromosome (or a set of variables) is chosen randomly, where each chromosome consists of n genes (or descriptors) represented by a single bit (step 1). Worth noting is that each variable (or descriptor) is treated by a binary coding (*i.e.*, the value 1 for the selected variable and 0 for the non-selected one) and the original population size is generally between 50 and 500. In step 2, the response is evaluated based on a criterion. Criteria include a maximum number of generations, the extent of improvement in the response, the number of iterations, the total calculation time and so on. When the GA is applied to a regression technique (as a fitness function), the response is generally the cross-validated (CV) variance (such as Q^2 values) explained by the selected variable combination (*i.e.*, CV explained variance [%]). Step 3 (reproduction) creates a new population (forming the next generation) with the surviving chromosomes through cross-over and mutation. Steps 2-3 are repeated until an evaluation criterion for termination is reached [65, 68, 69]. The final model can be obtained by a stepwise (forward or backward) approach. The backward stepwise selection approach (employed throughout the present study) adds variables

sequentially based on their frequency of selection [60]. The final model is selected from a plot of response (% CV explained variance or Q^2_{CV}) *vs.* the number of variables, where the beginning point of the plateau in the plot is the best model to be selected [65].

1.3.3. Data collection and durability

Considering that chromatographic retention data are typically used as response data for a QSRR modelling, care should be taken to acquire reliable and durable retention data. Accordingly, it is desirable to prepare samples using the same sample treatments and then to analyse them under the same conditions (such as mobile phase compositions, column and detection temperature, and flow-rate of mobile phase). 54 Systematic and efficient data collection is generally performed based on design of experiments (DoE) to handle a large size of samples (or data). Extensive databases embedded in the Virtual Column software (including the anion databases for QSRR modelling used in this project) were obtained according to an appropriate experimental design around 10 years ago and then were stored in a "column database" in the software [35]. As an example, a full factorial design consisting of 2 factors (total eluent concentrations [E_T] and the molar ratio of the doubly charged eluent [R]) and 3 levels (low, medium, and high) was used to obtain retention data for a dual-species eluent (*i.e.*, carbonate/bicarbonate) systems. In other words, the retention data were collected at nine eluent compositions for this system [35]. The databases (embedded in the Virtual Column software) cover over 150 analytes (such as anions, cations, and carbohydrate species), 2 column diameters (4 mm i.d. and 2 mm i.d.), 20 columns, 5 eluent types, and 3 temperatures [38].

When using these embedded data to predict retention times of analytes on recently produced columns, prediction errors using the Virtual Column software can be observed due to the possibility of changes in column behaviour arising primarily from loss of functional groups from the stationary phase leading to reduced ion-exchange

capacity [70]. To improve data durability and hence, to enable more reliable *in silico* optimization of IC separations on new versions of columns by the Virtual Column software, "*porting*" methodology has been developed to recalibrate and update the embedded retention data with minimal experimental input [70]. Briefly, a "porting equation", expressing the relationship between new and old (or embedded) data was derived using only two representative ions (chloride and thiosulfate). For this purpose, retention times of these two ions (*i.e.*, new data) were determined experimentally on a new column under three eluent concentrations. Finally, retention data for the entire analyte databases were updated by substituting old (or embedded) data into the porting equation. It was reported that average prediction errors less than 3% were observed for ported anionic and cationic analytes under various experimental conditions (such as column internal diameters and complex elution profiles) [70]. New "porting" methodology to improve the prediction errors on various columns will be discussed in Chapter 3.

1.3.4. Modelling techniques

When correlating the chromatographic retention parameters with the molecular descriptors in QSRR, two statistical techniques: multiple linear regression (MLR) and partial least square (PLS) regression have been commonly used [71-77]. MLR, based on a least-squares procedure, is the most popular approach in QSRR due to its simplicity and easy interpretation, and the general expression is as follows [71, 72, 78]:

$$y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_nX_n \quad \textbf{(1.17)}$$

where y is the dependent variable (or chromatographic retention parameter), $X_1, X_2, \cdots, X_n$ are independent variables (or descriptors), $a_1, a_2, \cdots, a_n$ are regression coefficients, a_0 is the intercept of the model, and n is the number of independent variables. The regression coefficients and their algebraic signs in the model equation represent independent contributions of each descriptor to the dependent variable [71]

and the statistical significance of the regression coefficients can be checked by the *p*-value (<0.05) from a '*t*' test [78]. In addition, other parameters, such as the determination coefficient R^2 , Fisher (or variance) ratio *F*, and standard error of estimation *s*, are also employed to evaluate the statistical significance of the generated MLR models [78]:

$$R^2 = 1 - \frac{\sum(y_{\text{exp}} - y_{\text{pred}})^2}{\sum(y_{\text{exp}} - \overline{y_{\text{exp}}})^2} \quad \textbf{(1.18)}$$

$$F = \frac{\frac{\sum(y_{\text{pred}} - \overline{y_{\text{exp}}})^2}{m}}{\frac{\sum(y_{\text{exp}} - y_{\text{pred}})^2}{n-m-1}} \quad \textbf{(1.19)}$$

$$s = \sqrt{\frac{(y_{\text{exp}} - y_{\text{pred}})^2}{n-m-1}} \quad \textbf{(1.20)}$$

where y_{exp} and y_{pred} are the experimental and predicted (response) values, $\overline{y_{\text{exp}}}$ is the mean of the experimental values, *n* the number of descriptors, *m* the number of objects (or compounds), respectively. The *m* and *n*-*m*-1 in **Eq. 1.19** and **1.20** indicate degrees of freedom. As R^2 values are closer to 1, the fitting quality of the models are better. The square root of R^2 is referred as to the multiple correlation coefficient *R*. The fitting quality of the models can be visualised from scatter plots of measured *vs*. predicted values. When the models are statistically more significant, the Fisher ratio *F* values are higher and the standard errors of estimate of *y* (*i.e.*, *s* values) low. Worth noting is that the *F* ratio is used as a measure of overall significance of the regression coefficients [78].

When handling a much larger number of descriptors for a smaller number of compounds, PLS is preferred over MLR and can be expressed as [73]:

$$y = a_1 LV_1 + a_2 LV_2 + a_3 LV_3 + \cdots + a_{\text{i}} LV_{\text{i}} + \cdots + a_{\text{m}} LV_{\text{m}} \quad \textbf{(1.21)}$$

where y is the dependent variable (or chromatographic retention parameters), $LV_1, LV_2, \cdots, LV_{\text{i}}, \cdots, LV_{\text{n}}$ are independent variables (or latent variables [LV]), $a_1, a_2, \cdots, a_{\text{i}}, \cdots, a_{\text{n}}$ are regression coefficients, and *m* is the number of LVs. The LVs are

a linear combination of descriptors **(Eq. 1.21)** and the ith latent variable LV_i can be expressed as a function of the original variables (*i.e.*, descriptors):

$$LV_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n \quad \textbf{(1.22)}$$

where $X_1, X_2, \cdots, X_n$ are independent variables (or descriptors), and *n* is the number of descriptors. The PLS extracts components (or *LV*s) sequentially from the descriptor matrix **X** that best predict the dependent variable *y*, maximising the covariance between the descriptor matrix and the retention parameters. The optimal number of *LV*s is generally determined by cross-validation where the addition of new components is no longer significant in predictions [78]. The PLS can minimise some limitations for the MLR, including chance correlation, over-fitting, and multi-collinearity of descriptors [60, 72-74].

1.3.5. Role of similarity

According to the "Similar Property Principle" [79], molecules having similar structures tend to have similar physicochemical properties. Accordingly, the concept of molecular similarity, along with a variety of similarity measures, has been extensively employed for the investigation of Quantitative Structure-Activity Relationships (QSARs), and the virtual screening of compounds in medicinal chemistry, and also for QSRR [32, 77, 80-83].

Sheridan *et al.* [81] has reported that the similarity of a target molecule (*i.e.*, a molecule about which a prediction is to be made) to molecules in the training set plays an important role in generating predictive QSAR models for the target molecule. Two main factors were important: first, the similarity between the target molecule and the most similar molecule in the training set, and second, the number of molecules in the training set having similarity values to the target molecule greater than a user-defined threshold value. In addition, this trend was more applicable for narrow training sets rather than diverse training sets.

Another QSAR study has reported that predictive QSAR models can be generated using similar clusters for a query compound [80]. Clusters comprised of similar compounds were obtained by hierarchical clustering analysis using a fathead minnow dataset containing 322 organic compounds, based on standardized Euclidean- and city-block similarity [80]. Briefly, the hierarchical clustering analysis, a popular technique using the similarity concept, is categorised into two types, based on the procedure of the cluster's creations: agglomerative (or a bottom–up method) and divisive (or a top–down method). In the divisive approach, the clusters are formed by successive splitting from a big cluster and the agglomerative approach follows the reverse process (*i.e.*, bigger clusters are created by combining smaller clusters, starting with each individual object) [80, 84]. The splitting or fusion of the clusters is performed by a linkage method which is based on the similarities (or distances) between clusters [80]. This technique is generally represented by a graph with a tree structure named a dendrogram and has advantages to alter easily similarity measurement criteria and linkage methods suitable for different applications [84].

Table 1.3 lists various molecular similarity measures based on 2-dimensional (or 2D) fingerprints, along with the Euclidean- and city-block similarity coefficients used in the QSAR study described above [82, 85]. The 2D fingerprint-based molecular similarity coefficient is a popular measure of the intermolecular structural similarity between two molecules and is determined based on binary strings encoding the existence (*i.e.*, presence or absence) of the structural fragments in the molecules [82, 86]. These similarity measures have been extensively used in drug discovery to search chemical structures in a large databases [82, 86-88]. Chen and Reynolds [85] have reported that the Tanimoto similarity coefficient outperformed the Euclidean distance in the effectiveness of similarity searching in retrieving active structural analogues. Indeed, the Tanimoto coefficient, the most popular 2D fingerprint-based similarity measure, has been the most commonly used in both in-house and commercial software

Table 1.3 Similarity coefficients for use with 2D fingerprints [82, 85, 86].

Similarity and distance coefficient	Expression
Tanimoto	$S_{\mathrm{AB}} = \dfrac{c}{a+b-c}$
Euclidean	$D_{\mathrm{AB}} = \sqrt{a+b-2c}$
City-block (or Hamming or Manhattan)	$D_{\mathrm{AB}} = a+b-2c$
Cosine	$S_{\mathrm{AB}} = \dfrac{c}{\sqrt{ab}}$
Russell-Rao	$S_{\mathrm{AB}} = \dfrac{c}{n_{tot}}$
Forbes	$S_{\mathrm{AB}} = \dfrac{cn_{tot}}{ab}$

For two molecules, A and B, represented by binary fingerprints, parameters in equations imply: ***a***: the number of unique fragments (or bits) in molecule A; ***b***: the number of unique fragments in molecule B; ***c***: the number of unique fragments in both of molecules A and B; $\boldsymbol{n_{tot}}$: a total number of bits in both of molecules A and B. ***S*** and ***D*** denote similarities and distances, respectively.

packages, such as ChromGenius [89]. In addition, the Tanimoto coefficient has a range of zero (no common bits) to one (all bits the same) [82]. The Tanimoto similarity coefficient was therefore employed in the present studies for similarity searching, ranking a database by decreasing order of similarity, to include similar compounds into the training sets for generating QSRR models (Chapters 5 and 6). It should be noted that Hamming and Euclidean coefficients in **Table 1.3** represent the distances of two molecules and their similarity values can be transformed by [90]:

$$S_{AB} = \frac{1}{1+ D_{AB}} \qquad (0 \leq S_{AB} \leq 1) \qquad \textbf{(1.23)}$$

Larger values of distance (*i.e.*, higher degree of dissimilarity) between two molecules result in lower values of similarity between those two molecules.

The Tanimoto similarity (TS) coefficient has previously been used to identify clusters comprised of similar compounds to a target compound, followed by the generation of QSRR models, to improve the predictive power of the developed models [83]. The 20 most similar compounds to a target compound were ranked from a dataset of 86 suspected doping compounds, based on their TS score and were then included in the training set for constructing a GA-PLS model for the target compound. By this approach called "federation of local models", the prediction accuracy of local models was improved significantly and this approach was employed in the present study (Chapters 5 and 6).

1.4. Aims of this study

The ultimate objective of this work was to develop predictive Quantitative Structure-Retention relationship (QSRR) models enabling rapid "scoping" method development in IC. Using the QSRR models, predictions of retention times of target analytes based solely on the analytes' structures can be employed to screen potential stationary phases for successful IC separations of the target analytes without

experimentation, which can lead to reductions in the cost and the time in IC method development.

The first aim was to update retention data of anions embedded in the Virtual Column software by a modified "porting" methodology, improving the retention predictions on a wide range of newly produced columns. The recalibration of the embedded retention data on new columns is important since reliable and robust retention data are an essential factor for successful QSRR modelling and the QSRR models can be used in practice to predict retention times of analytes on recently produced new columns.

The second aim was to generate QSRR models for *a*- and *b*-values in the LSS model in IC by evolutionary algorithm-multiple linear regression (EA-MLR). The great advantage of QSRR modelling for *a*- and *b*-values of analytes, instead of directly predicting retention times of target analytes, is that the predicted *a*- and *b*-values allow the predictions of retention times of the analytes for all elution concentrations under isocratic- and gradient- elution modes, as well as multi-step eluent profiles comprised of sequential isocratic and linear gradient eluent conditions.

The final aim was to improve the predictive ability of QSRR models for new target analytes, not involved in the model generation, by introducing the concept of similarity screening into QSRR modelling. Two types of similarity measures (structural- and chromatographic- similarity indices) to construct local QSRR models for each target ion in the anions and cation database were to be investigated to create clusters of similar ions (compared to each target ion) into the training sets prior to QSRR modelling using GA-PLS regression. An investigation for various similarity measures (Tanimoto similarity index, *k*-ratio-, dual- and log*P*-dual-fliter) in QSRR modelling was required to identify the most suitable similarity measure reflecting IC retention, leading to the development of predictive QSRR models.

1.5. References

[1] P.R. Haddad, P.E. Jackson, Ion chromatography : principles and applications, in: Journal of Chromatography Library, Elsevier, Amsterdam, The Netherlands, 1990.

[2] P.R. Haddad, Ion chromatography, Anal. Bioanal. Chem. 379 (2004) 341-343.

[3] J. Weiss, Ion chromatography, 2nd ed., Wiley-VCH, Weinheim, 1995.

[4] J.S. Fritz, D.T. Gjerde, Ion chromatography, Wiley-VCH, Weinheim, 2009.

[5] P.R. Haddad, Developments in detection methods for ion chromatography, Chromatographia 24 (1987) 217-225.

[6] J. Weiss, D. Jensen, Modern stationary phases for ion chromatography, Anal. Bioanal. Chem. 375 (2003) 81-98.

[7] S. Rabin, J. Stillian, Practical aspects on the use of organic solvents in ion chromatography, J. Chromatogr. A 671 (1994) 63-71.

[8] C.A. Pohl, Recent developments in ion-exchange columns for ion chromatography, LCGC North America 31 (2013) 16-22.

[9] D. Jensen, J. Weiss, M.A. Rey, C.A. Pohl, Novel weak acid cation-exchange column J. Chromatogr. 640 (1993) 65-71.

[10] C.A. Pohl, C. Saini, New developments in the preparation of anion exchange media based on hyperbranched condensation polymers, J. Chromatogr. A 1213 (2008) 37-44.

[11] R.W. Slingsby, C.A. Pohl, Anion-exchange selectivity in latex-based columns for ion chromatography, J. Chromatogr. 458 (1988) 241-253.

[12] R.E. Barron, J.S. Fritz, Effect of functional group structure on the selectivity of low-capacity anion exchangers for monovalent anions, J. Chromatogr. 284 (1984) 13-25.

[13] C. Liang, C.A. Lucy, Characterization of ion chromatography columns based on hydrophobicity and hydroxide eluent strength, J. Chromatogr. A 1217 (2010) 8154-8160.

[14] N. Karu, J.P. Hutchinson, G.W. Dicinoski, M. Hanna-Brown, K. Srinivasan, C.A. Pohl, P.R. Haddad, Determination of pharmaceutically related compounds by suppressed ion chromatography: IV. Interfacing ion chromatography with universal detectors, J. Chromatogr. A 1253 (2012) 44-51.

[15] P.J. Dumont, J.S. Fritz, L.W. Schmidt, Cation-exchange chromatography in non-aqueous solvents, J. Chromatogr. A 706 (1995) 109-114.

[16] P. Zakaria, G.W. Dicinoski, B.K. Ng, R.A. Shellie, M. Hanna-Brown, P.R. Haddad, Application of retention modelling to the simulation of separation of organic anions in suppressed ion chromatography, J. Chromatogr. A 1216 (2009) 6600-6610.

[17] P. Zakaria, G. Dicinoski, M. Hanna-Brown, P.R. Haddad, Prediction of the effects of methanol and competing ion concentration on retention in the ion chromatographic separation of anionic and cationic pharmaceutically related compounds, J. Chromatogr. A 1217 (2010) 6069-6076.

[18] X. Ding, S. Mou, Ion chromatographic analysis of tetracyclines using polymeric column and acidic eluent, J. Chromatogr. A 897 (2000) 205–214.

[19] N. Karu, G.W. Dicinoski, P.R. Haddad, Use of suppressors for signal enhancement of weakly-acidic analytes in ion chromatography with universal detection methods, TrAC 40 (2012) 119-132.

[20] L.R. Snyder, J.W. Dolan, J.R. Gant, Gradient elution in high-performance liquid chromatography. I. Theoretical basis for reversed-phase systems, J. Chromatogr. 165 (1979) 3-30.

[21] M. Gilar, H. Xie, A. Jaworski, Utility of retention prediction model for investigation of peptide separation selectivity in reversed-phase liquid chromatography: Impact of concentration of trifluoroacetic acid, column temperature, gradient slope and type of stationary phase, Anal. Chem. 82 (2010) 265–275.

[22] P.R. Haddad, P.E. Jackson, M.J. Shaw, Developments in suppressor technology for inorganic ion analysis by ion chromatography using conductivity detection, J. Chromatogr. A 1000 (2003) 725-742.

[23] H. Small, T.S. Stevens, W.C. Bauman, Novel ion exchange chromatographic method using conductimetric detection, Anal. Chem. 47 (1975) 1801-1809.

[24] W.W. Buchberger, Detection techniques in ion analysis: what are our choices?, J. Chromatogr. A 884 (2000) 3–22.

[25] N. Karu, G.W. Dicinoski, M. Hanna-Brown, K. Srinivasan, C.A. Pohl, P.R. Haddad, Determination of pharmaceutically related compounds by suppressed ion chromatography: III. Role of electrolytic suppressor design, J. Chromatogr. A 1233 (2012) 71-77.

[26] S. Mallipattu, G. Sévenier, Y. Dudal, A new HPIC-UV method to quantify phenolic acids in food and environmental samples, Int. J. Environ. Anal. Chem. 90 (2010) 633-643.

[27] M.D. Shastri, C. Johns, J.P. Hutchinson, M. Khandagale, R.P. Patel, Ion exchange chromatographic separation and isolation of oligosaccharides of intact low-molecular-weight heparin for the determination of their anticoagulant and anti-inflammatory properties, Anal. Bioanal. Chem. 405 (2013) 6043-6052.

[28] R. Ghanem, M.A. Bello, M.C.A. Guiraum, Determination of beta-blocker drugs in pharmaceutical preparations by non-suppressed ion chromatography, J. Pharm. Biomed. Anal. 15 (1996) 383-388.

[29] H. Liu, H. Wang, V.B. Sunderland, An isocratic ion exchange HPLC method for the simultaneous determination of flucloxacillin and amoxicillin in a pharmaceutical formulation for injection, J. Pharm. Biomed. Anal. 37 (2005) 395-398.

[30] J.L. Perez, M.A. Bello, Determination of paracetamol in dosage forms by non-suppressed ion chromatography, Talanta 48 (1999) 1199–1202.

[31] J. Weiss, Ion chromatography - A review of recent developments, Fresenius Z Anal. Chem. 327 (1987) 451-455.

[32] E. Tyteca, M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad, Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Models for retention prediction: use of retention factor ratio, J. Chromatogr. A 1486 (2017) 50-58.

[33] K.M. Biswas, B.C. Castle, B.A. Olsen, D.S. Risley, M.J. Skibic, P.B. Wright, A simple and efficient approach to reversed-phase HPLC method screening, J. Pharm. Biomed. Anal. 49 (2009) 692-701.

[34] T. Baczek, R. Kaliszan, H.A. Claessens, M.A.v. Straten, Computer-assisted optimisation of reversed-phase HPLC isocratic separations of neutral compounds, LCGC Europe 14 (2001) 304-313.

[35] J.E. Madden, M.J. Shaw, G.W. Dicinoski, N. Avdalovic, P.R. Haddad, Simulation and optimization of retention in ion chromatography using virtual column 2 software, Anal. Chem. 74 (2002) 6023-6030.

[36] I. Molnar, Computerised design of separation strategies by reversed-phase liquid chromatography: development of DryLab software, J. Chromatogr. A 965 (2002) 175–194.

[37] P. Zakaria, G. Dicinoski, M. Hanna-Brown, P.R. Haddad, Modelling and optimisation of ion chromatographic separations of pharmaceutically relevant organic ions, in: L. Bhattacharyya, J.S. Rohrer (Eds.), Applications of ion chromatography for

pharmaceutical and biological products Chapter 5 Modelling and optimisation of ion chromatographic separations of pharmaceutically relevant organic ions, John Wiley & Sons, Inc., New Jersey, 2012, pp. 107-133.

[38] B.K. Ng, T.T.Y. Tan, R.A. Shellie, G.W. Dicinoski, P.R. Haddad, Computer-assisted simulation and optimisation of retention in ion chromatography, TrAC 80 (2016) 625-635.

[39] E. Tyteca, S.H. Park, R.A. Shellie, P.R. Haddad, G. Desmet, Computer-assisted multi-segment gradient optimization in ion chromatography, J. Chromatogr. A 1381 (2015) 101-109.

[40] R.A. Shellie, B.K. Ng, G.W. Dicinoski, S.D. Poynter, J.W. O'Reilly, C.A. Pohl, P.R. Haddad, Prediction of analyte retention for ion chromatography separations performed using elution profiles comprising multiple isocratic and gradient steps, Anal. Chem. 80 (2008) 2474-2482.

[41] R.D. Rocklin, C.A. Pohl, J.A. Schibler, Gradient elution in ion chromatography, J. Chromatogr. 411 (1987) 107-119.

[42] P. Jandera, J. Churacek, Gradient elution in liquid chromatography I. The influence of the composition of the mobile phase on the capacity ratio (retention volume, band width, and resolution) in isocratic elution - theoretical considerations, J. Chromatogr. 91 (1974) 207-221.

[43] E. Van Gyseghem, S. Van Hemelryck, M. Daszykowski, F. Questier, D.L. Massart, Y. Vander Heyden, Determining orthogonal chromatographic systems prior to the development of methods to characterise impurities in drug substances, J. Chromatogr. A 988 (2003) 77-93.

[44] K.P. Xiao, Y. Xiong, F.Z. Liu, A.M. Rustum, Efficient method development strategy for challenging separation of pharmaceutical molecules using advanced chromatographic technologies, J. Chromatogr. A 1163 (2007) 145-156.

[45] R.M. Krisko, K. McLaughlin, M.J. Koenigbauer, C.E. Lunte, Application of a column selection system and DryLab software for high-performance liquid chromatography method development, J. Chromatogr. A 1122 (2006) 186-193.

[46] W-D. Beinert, V. Eckert, S. Galushko, V. Tanchuk, I. Shishkina, Automated HPLC method development: A step forward with innovative software technology, LCGC Europe On-line supplement, (2001) 34-38.

[47] L.R. Snyder, J.W. Dolan, P.W. Carr, The hydrophobic-subtraction model of reversed-phase column selectivity, J. Chromatogr. A 1060 (2004) 77-116.

[48] S.V. Galushko, Calculation of retention and selectivity in reversed-phase liquid chromatography, J. Chromatogr. 552 (1991) 91-102.

[49] S.V. Galushko, The calculation of retention and selectivity in reversed-phase liquid chromatography II. Methanol-water eluents, Chromatographia 36 (1993) 39-42.

[50] R. Kaliszan, QSRR: Quantitative Structure-(Chromatographic) Retention Relationships, Chem. Rev. 107 (2007) 3212-3246.

[51] K. Heberger, Quantitative structure-(chromatographic) retention relationships, J. Chromatogr. A 1158 (2007) 273-305.

[52] R. Kaliszan, Quantitative structure-retention relationships, Anal. Chem. 64 (1992) 619A-631A.

[53] R. Todeschini, V. Consonni, Handbook of molecular descriptors, Wiley-VCH, Weinheim, 2000.

[54] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure–retention relationship studies, J. Chromatogr. A 988 (2003) 261-276.

[55] R. Kaliszan, Quantitative structure-retention relationships (QSRR) in chromatography, in: I.D. Wilson (Ed.) Encyclopedia of separation science, Academic Press, San Diego, 2000, pp. 4063-4075.

[56] Dragon 6.0, Talete, Milano, Italy, 2014, talete.mi.it, in.

[57] A.M. Helguera, R.D. Combes, M.P. González, M.N.D.S. Cordeiro, Applications of 2D Descriptors in Drug Design: A DRAGON Tale, Curr. Top. Med. Chem. 8 (2008) 1628-1655.

[58] V. Rastija, M. Medić-Šarić, QSAR modeling of anthocyanins, anthocyanidins and catechins as inhibitors of lipid peroxidation using three-dimensional descriptors, Med. Chem. Res. 18 (2008) 579-588.

[59] M. Goodarzi, R. Jensen, Y. Vander Heyden, QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions, J. Chromatogr. B 910 (2012) 84-94.

[60] M. Talebi, G. Schuster, R.A. Shellie, R. Szucs, P.R. Haddad, Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography, J. Chromatogr. A 1424 (2015) 69-76.

[61] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemom. Intell. Lab. Syst. 76 (2005) 185-196.

[62] P. Zuvela, J.J. Liu, K. Macur, T. Baczek, Molecular descriptor subset selection in theoretical peptide quantitative structure-retention relationship model development using nature-inspired optimization algorithms, Anal. Chem. 87 (2015) 9876-9883.

[63] Š. Ukić, M. Novak, A. Vlahović, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part II: Artificial intelligence QSRR approach, Chromatographia 77 (2014) 997-1007.

[64] G. Carlucci, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure-retention relationships, Anal. Chim. Acta 601 (2007) 68-76.

[65] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.

[66] J.H. Holland, Adaptation in natural and artificial systems, MIT Press, Cambridge, MA, 1992.

[67] G. Jones, Genetic and evolutionary algorithms, in: P.v.R. Schleyer (Ed.) Encyclopedia of computational chemistry, John Wiley and Sons, 1998, pp. 1115-1127.

[68] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemom. 6 (1992) 267-281.

[69] R. Leardi, Genetic algorithms in chemistry, J. Chromatogr. A 1158 (2007) 226-233.

[70] B.K. Ng, R.A. Shellie, G.W. Dicinoski, C. Bloomfield, Y. Liu, C.A. Pohl, P.R. Haddad, Methodology for porting retention prediction data from old to new columns and from conventional-scale to miniaturised ion chromatography systems, J. Chromatogr. A 1218 (2011) 5512-5519.

[71] J. Ghasemi, S. Saaidpour, QSRR Prediction of the Chromatographic Retention Behavior of Painkiller Drugs, J. Chromatogr. Sci. 47 (2009) 156-163.

[72] V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, Prediction of capillary gas chromatographic retention times of fatty acid methyl

esters in human blood using MLR, PLS and back-propagation artificial neural networks, Talanta 83 (2011) 1014-1022.

[73] C.B. Mazza, C.E. Whitehead, C.M. Brenernan, S.M. Crarner, Predictive quantitative structure retention relationship models for ion-exchange chromatography, Chromatographia 56 (2002) 147-152.

[74] Š. Ukić, M. Novak, P. Žuvela, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part I: Conventional QSRR approach, Chromatographia 77 (2014) 985-996.

[75] K. Gorynski, B. Bojko, A. Nowaczyk, A. Bucinski, J. Pawliszyn, R. Kaliszan, Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds, Anal. Chim. Acta 797 (2013) 13-19.

[76] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures, J. Chromatogr. A 1486 (2017) 59-67.

[77] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A. Pohl, Rapid method development in hydrophilic interaction liquid chromatography for pharmaceutical analysis using a combination of quantitative structure-retention relationships and design of experiments, Anal. Chem. 89 (2017) 1870-1878.

[78] K. Roy, S. Kar, R.N. Das, Statistical methods in QSAR/QSPR, in: A primer on QSAR/QSPR modelling, Springer International Publishing, 2015, pp. 37-59.

[79] M.A. Johnson, G.M. Maggiora, Concepts and applications of molecular similarity, John Wiley & Sons, New York, 1990.

[80] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions, J. Mol. Graph. Model. 23 (2005) 503-523.

[81] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, S.K. Kearsley, Similarity to molecules in the training set Is a good discriminator for prediction accuracy in QSAR, J. Chem. Inf. Comput. Sci. 44 (2004) 1912-1928.

[82] P. Willett, Similarity-based virtual screening using 2D fingerprints, Drug Discovery Today 11 (2006) 1046-1053.

[83] M. Talebi, S.H. Park, M. Taraji, Y. Wen, R.I.J. Amos, P.R. Haddad, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan, Retention time prediction based on molecular structure in pharmaceutical method development: A perspective, LCGC North America 34 (2016) 550-558.

[84] C. Tistaert, B. Dejaegher, Y. Vander Heyden, Chromatographic separation techniques and data handling methods for herbal fingerprints: a review, Anal. Chim. Acta 690 (2011) 148-161.

[85] X. Chen, C.H. Reynolds, Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients, J. Chem. Inf. Comput. Sci. 42 (2002) 1407-1414.

[86] P. Willett, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983-996.

[87] S. Arif, N.Z.S. Khan, N. Malim, S. Zainudin, Retrieval performance using different type of similarity coefficient for virtual screening, Res. J. Appl. Sci. Eng. Tech. 9 (2015) 391-395.

[88] P. Willet, V. Winterman, Implementation of nearest-neighbor searching in an online chemical structure search system, J. Chem. Inf. Comput. Sci. 26 (1986) 36-41.

[89] Y.V. Kazakevich, R. LoBrutto, HPLC for Pharmaceutical Scientists, John Wiley & Sons, Inc., New Jersey, 2007.

[90] R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scaled functions, Chemom. Intell. Lab. Syst. 87 (2007) 3-17.

Chapter 2

Experimental

2.1. Reagents

Only chemicals of analytical reagent grade were used in this study. All chemicals used are listed in **Table 2.1**.

2.2. Preparation of standard solutions

Standard solutions were prepared by dissolving the salts listed in **Table 2.1** in Milli-Q water (18.2 MΩ; Merck-Millipore, Bayswater, Australia). Stock solutions with concentrations of 1000 mg/L (based on the mass of either anion or cation) were diluted to prepare working solutions for injections within the concentrations range of 5 - 100 mg/L.

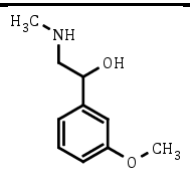
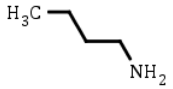
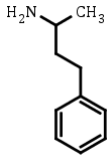
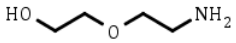
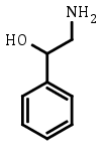
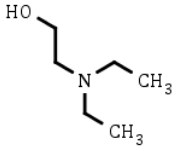
2.3. Columns

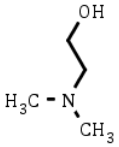
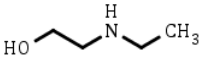
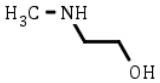
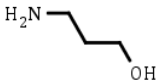
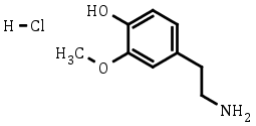
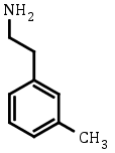
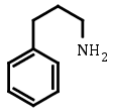
Three anion-exchange columns (Thermo Scientific™ Dionex™ IonPac™ AS20, AS19, and AS11HC) and one cation-exchange column (Thermo Scientific™ Dionex™ IonPac™ CS17) were employed in this study and their characteristics are given in **Table 2.2** [1].

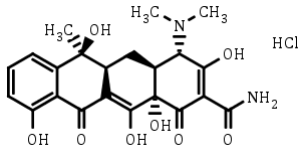
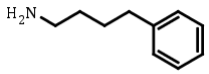
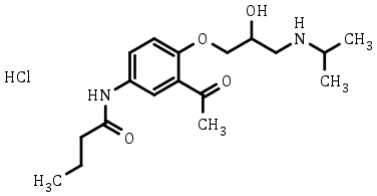
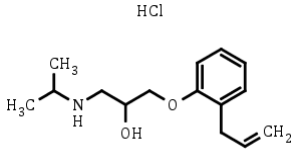
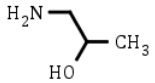
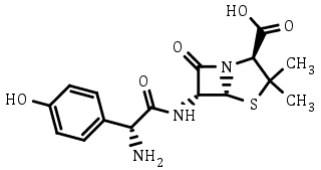
2.4. Instrumentation

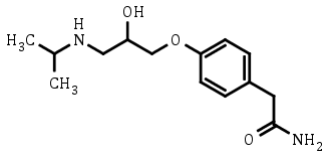
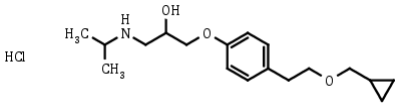
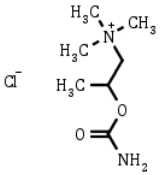
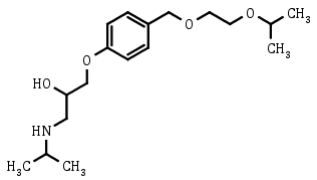
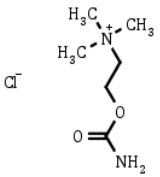
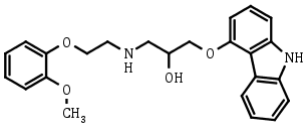
The IC system used in this study was a Dionex (Sunnyvale, CA, USA) ICS-3000™, consisting of a dual gradient pump unit (Dionex ICS-3000 DP), dual eluent generator unit (Dionex ICS-3000 EG), dual suppressed conductivity detector compartment (Dionex ICS-3000 DC) and autosampler (Dionex AS). A Dionex EluGen® cartridge (EGC II KOH, and EGC II Methanesulfonic acid (MSA), for anionic and cationic analysis,

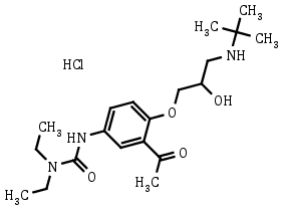
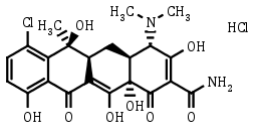
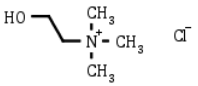
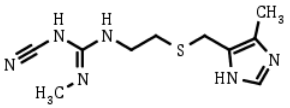
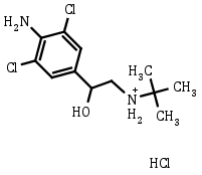
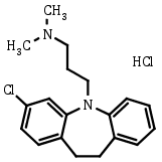
Table 2.1 Chemicals utilized in this study

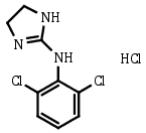
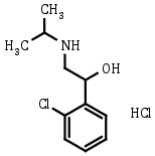
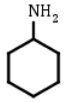
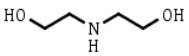
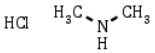
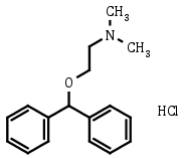
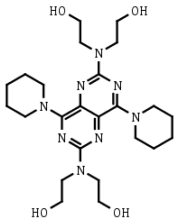
Compound	Structure	CAS Number	Supplier
1,3-Methoxyphenyl-2-methylaminoethanol		92188-49-3	Santa Cruz Biotechnology
1-Butylamine		109-73-9	Santa Cruz Biotechnology
1-Methyl-3-phenylpropylamine		22374-89-6	Fluka
2,2-Aminoethoxyethanol		929-06-6	Santa Cruz Biotechnology
2-Amino-1-phenylethanol		7568-93-6	Sigma-Aldrich
2-(Diethylamino)ethanol		100-37-8	Sigma-Aldrich

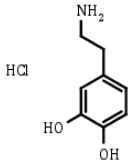
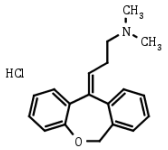
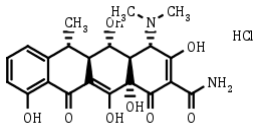
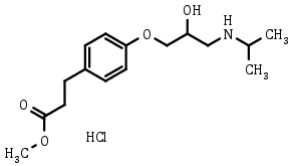
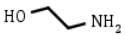
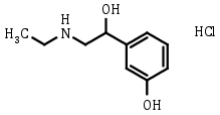
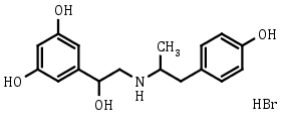
2-Dimethylaminoethanol		108-01-0	Santa Cruz Biotechnology
2-(Ethylamino)ethanol		110-73-6	Santa Cruz Biotechnology
2-(Methylamino)ethanol		109-83-1	Sigma-Aldrich
3-Amino-1-propanol		156-87-6	Santa Cruz Biotechnology
3-Methoxytyramine hydrochloride		1477-68-5	Santa Cruz Biotechnology
3-Methylphenethylamine		55755-17-4	Sigma-Aldrich
3-Phenylpropylamine		2038-57-5	Fluka

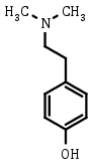
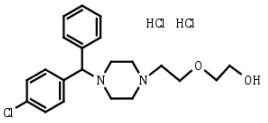
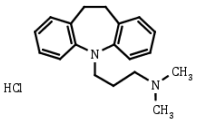
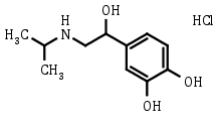
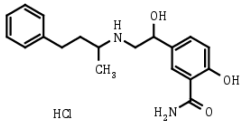
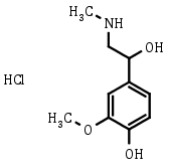
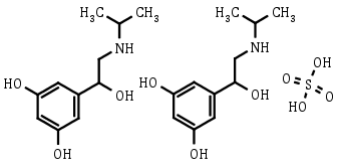
4-Epitetracycline hydrochloride		23313-80-6	Santa Cruz Biotechnology
4-Phenylbutylamine		13214-66-9	Aldrich
5-Amino-1-pentanol		2508-29-4	Santa Cruz Biotechnology
Acebutolol hydrochloride		34381-68-5	Santa Cruz Biotechnology
Alprenolol hydrochloride		13707-88-5	Santa Cruz Biotechnology
Amino-2-propanol		78-96-6	Santa Cruz Biotechnology
Amoxicillin		26787-78-0	Sigma

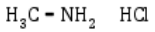
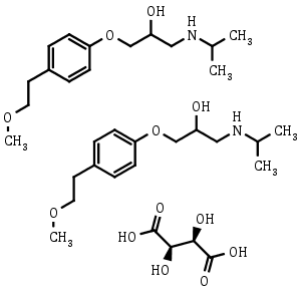
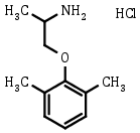
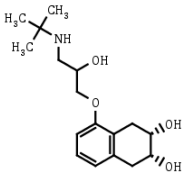
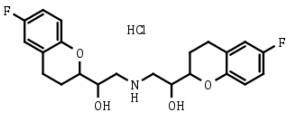
Atenolol		29122-68-7	Sigma-Aldrich
Betaxolol hydrochloride		63659-19-8	Sigma
Bethanechol chloride		590-63-6	Sigma-Aldrich
Bisoprolol		66722-44-9	Sigma-Aldrich
Carbachol		51-83-2	Santa Cruz Biotechnology
Carvedilol		72956-09-3	Santa Cruz Biotechnology

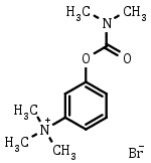
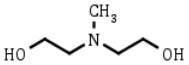
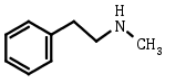
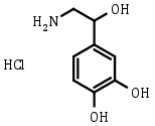
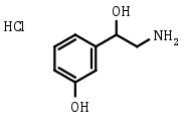
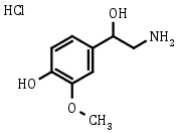
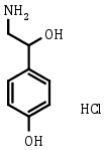
Celiprolol hydrochloride		57470-78-7	Santa Cruz Biotechnology
Chlortetracycline hydrochloride		64-72-2	Santa Cruz Biotechnology
Choline chloride		67-48-1	Sigma
Cimetidine		51481-61-9	Sigma
Clenbuterol hydrochloride		21898-19-1	Santa Cruz Biotechnology
Clomipramine hydrochloride		17321-77-6	Sigma

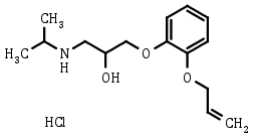
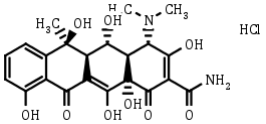
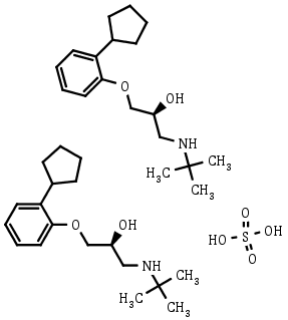
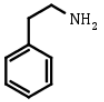
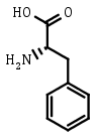
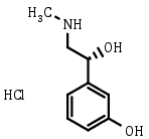
Clonidine hydrochloride		4205-91-8	Sigma
Clorprenaline hydrochloride		6933-90-0	Santa Cruz Biotechnology
Cyclohexylamine		108-91-8	Sigma-Aldrich
Diethanolamine		111-42-2	Fluka
Dimethylamine hydrochloride		506-59-2	Sigma
Diphenhydramine hydrochloride		147-24-0	Sigma
Dipyridamole		58-32-2	Sigma-Aldrich

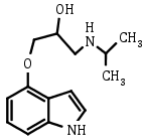
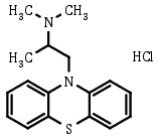
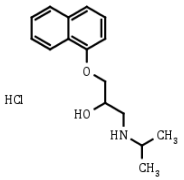
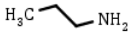
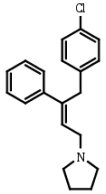
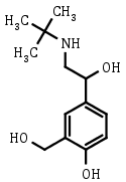
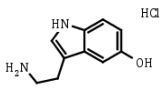
Dopamine hydrochloride		62-31-7	Sigma
Doxepin hydrochloride		1229-29-4	Sigma
Doxycycline hydrochloride		10592-13-9	Sigma
Esmolol hydrochloride		81161-17-3	Santa Cruz Biotechnology
Ethanolamine		141-43-5	Sigma
Etilefrine hydrochloride		943-17-9	Santa Cruz Biotechnology
Fenoterol hydrobromide		1944-12-3	Santa Cruz Biotechnology

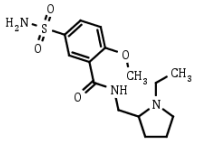
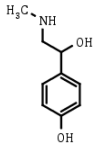
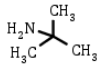
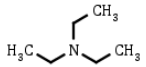
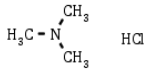
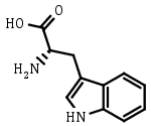
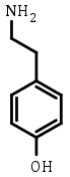
Hordenine		539-15-1	Sigma-Aldrich
Hydroxyzine dihydrochloride		2192-20-3	Sigma
Imipramine hydrochloride		113-52-0	Sigma
Isoprenaline hydrochloride		51-30-9	Sigma
Labetalol hydrochloride		32780-64-6	Sigma-Aldrich
DL-Metanephrine hydrochloride		881-95-8	Sigma-Aldrich
Metaproterenol hemisulfate		5874-97-5	Sigma

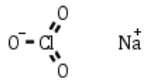
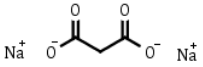
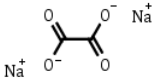
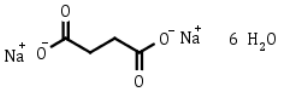
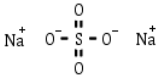
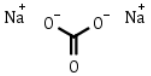
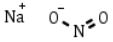
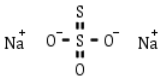
Methylamine hydrochloride		593-51-1	Sigma-Aldrich
(±)-Metoprolol (+)-tartrate salt		56392-17-7	Sigma-Aldrich
Mexiletine hydrochloride		31828-71-4	Sigma
Morpholine		110-91-8	Sigma-Aldrich
Nadolol		42200-33-9	Sigma-Aldrich
Nebivolol hydrochloride		152520-56-4	Sigma

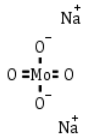
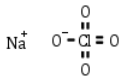
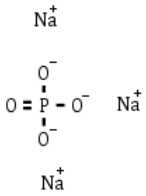
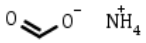
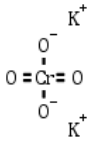
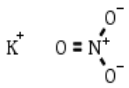
Neostigmine bromide		114-80-7	Sigma
N-Methyldiethanolamine		105-59-9	Sigma-Aldrich
N-Methylphenethylamine		589-08-2	Santa Cruz Biotechnology
N-Methylpyrrolidine		120-94-5	Sigma-Aldrich
DL-Norepinephrine hydrochloride		55-27-6	Sigma-Aldrich
Norfenefrine hydrochloride		4779-94-6	Sigma-Aldrich
DL-Normetanephrine hydrochloride		1011-74-1	Sigma-Aldrich
(±)-Octopamine hydrochloride		770-05-8	Sigma-Aldrich

Oxprenolol hydrochloride		6452-73-9	Santa Cruz Biotechnology
Oxytetracycline hydrochloride		2058-46-0	Santa Cruz Biotechnology
(S)-Penbutolol sulfate		38363-32-5	Santa Cruz Biotechnology
Phenethylamine		60-04-0	Santa Cruz Biotechnology
DL-Phenylalanine		150-30-1	Sigma
(R)-(-)-Phenylephrine hydrochloride		61-76-7	Santa Cruz Biotechnology

Pindolol		13523-86-9	Sigma
Promethazine hydrochloride		58-33-3	Sigma
Propranolol hydrochloride		318-98-9	Sigma
Propylamine		107-10-8	Fluka
Pyrrobutamine		91-82-7	Santa Cruz Biotechnology
Salbutamol		18559-94-9	Santa Cruz Biotechnology
Serotonin hydrochloride		153-98-0	Sigma

(S)-(-)-Sulpiride		23672-07-3	Sigma-Aldrich
(±)-Synephrine		94-07-5	Santa Cruz Biotechnology
tert-Butylamine		75-64-9	Santa Cruz Biotechnology
Triethylamine		121-44-8	Fluka
Trimethylamine hydrochloride		593-81-7	Fluka
DL-Tryptophan		54-12-6	Sigma
Tyramine		51-67-2	Sigma-Aldrich
Sodium fluoride		7681-49-4	BDH

Sodium chlorate		7775-09-9	BDH
Sodium malonate		141-95-7	BDH
Sodium oxalate		62-76-0	BDH
Sodium succinate hexahydrate		6106-21-4	BDH
Sodium sulfate		7757-82-6	BDH
Sodium carbonate		497-19-8	AJAX
Sodium nitrite		7632-00-0	AJAX
Sodium iodide		7681-82-5	Aldrich
Sodium thiosulfate		7772-98-7	Aldrich
Sodium bromide		7647-15-6	Sigma

Sodium molybdate		7631-95-0	Sigma
Sodium perchlorate		7601-89-0	Sigma
Sodium phosphate		7601-54-9	Sigma
Ammonium chloride		235-186-4	BDH
Ammonium formate		540-69-2	Sigma
Potassium chromate		7789-00-6	BDH
Potassium nitrate		7757-79-1	BDH

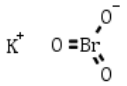
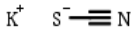
Potassium bromate		7758-01-2	AJAX
Potassium thiocyanate		333-20-0	AJAX

Table 2.2 Properties of ion-exchange columns [1, 2].

Column	**Length** (mm)	**Diameter** (mm)	**Ion exchange group**	**Hydrophobicity**	**Capacity** (μequiv)
AS 20	250	4	Alkanol quaternary ammonium ion	Ultralow	310
AS19	250	4	Alkanol quaternary ammonium ion	Low	240
AS11-HC	250	4	Alkanol quaternary ammonium ion Latex cross linking: 6% Latex diameter: 70 nm	Medium low	290
CS 17	250	2	Grafted carboxylic acid	Very low	363

respectively) were employed to electrolytically generate the eluent, followed by a Dionex CR-TC ion trap column and degasser. Chromeleon® chromatography management software (version 6.80) was used for instrument control and data acquisition. The column temperature of 30°C was used for both anionic and cationic analysis. Anions were analysed by conductivity detection at 35°C after eluent suppression using a Dionex ASRS 300 4 mm suppressor. More details are provided in Chapter 3. Cations were analysed by either non-suppressed UV detection or suppressed conductivity detection using a Dionex CSRS 300 2 mm suppressor, as per the system configurations shown in **Figure 2.1**. In addition, an organic solvent (36% (v/v) acetonitrile) stream was introduced after the CR-TC ion trap and then mixed with the MSA eluent through a T-piece connector followed by a gradient mixer (Dionex GM-4 2mm) [3]. The details for separation conditions, along with the other specifications such as detector type and other individual changes, are discussed in the relevant chapters.

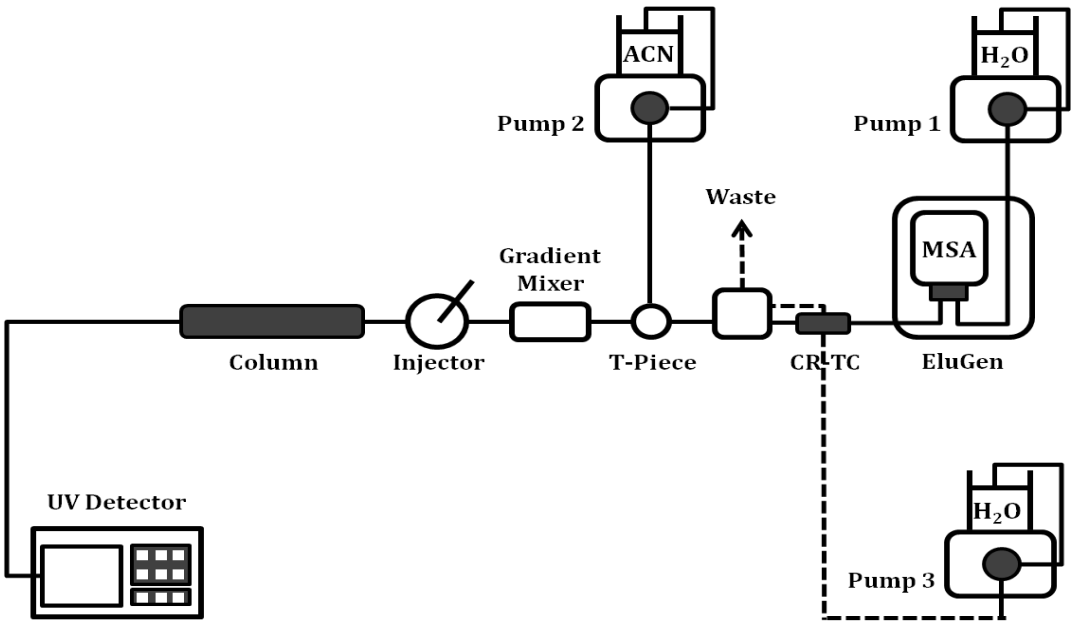
2.5. References

[1] Dionex Corporation, Sunnyvale, CA, https://www.thermofisher.com/au/en/home/industrial/chromatography/ion-chromatography-ic/ion-chromatography-columns.html [Accessed: January 2017].

[2] R.A. Shellie, E. Tyrrell, C.A. Pohl, P.R. Haddad, Column selection for comprehensive multidimensional ion chromatography, J. Sep. Sci. 31 (2008) 3287-3296.

[3] P. Zakaria, G.W. Dicinoski, B.K. Ng, R.A. Shellie, M. Hanna-Brown, P.R. Haddad, Application of retention modelling to the simulation of separation of organic anions in suppressed ion chromatography, J. Chromatogr. A 1216 (2009) 6600-6610.

a)



b)

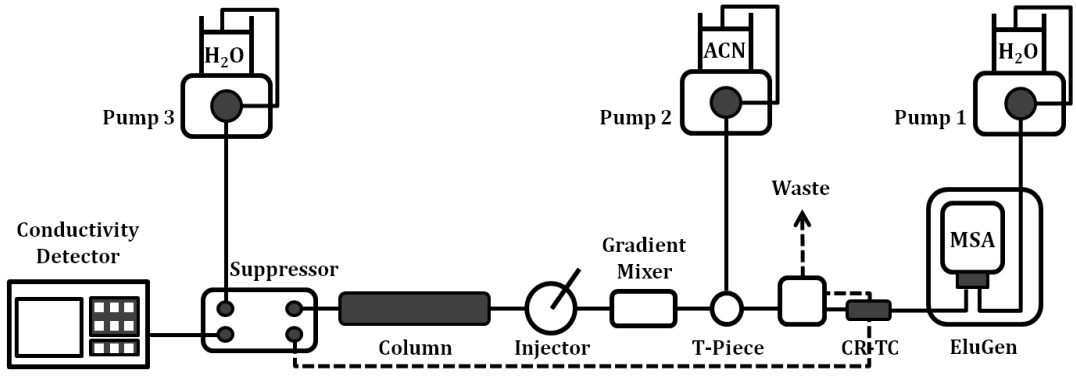


Figure 2.1 Schematic diagrams of experimental set-ups for a) non-suppressed UV detection and b) suppressed conductivity detection. Neat ACN is pumped using pump 2. Water pumped using pump 3 is to regenerate the suppressor.

Chapter 3

Porting methodology for the review of retention data

3.1. Introduction

Method translation or method transfer in chromatographic analysis has become an area of increasing interest. Numerous studies regarding method transfer have been reported in the area of liquid chromatography (LC) [1, 2] as well as gas chromatography (GC) [3, 4]. Method translation in GC is described as the rescaling of method parameters (temperature programs, pressures, *etc.*) as well as GC components (carrier gases, columns, detectors, *etc.*) without losing the peak elution pattern [3]. This can lead to the improvement of chromatographic performance in areas such as sample capacity, analysis time, and peak resolution through simple rescaling, along with the reduction of the development time and cost required for the creation of a desired chromatographic analysis method.

Recently, the concept of method transfer has been introduced to update the extensive retention databases embedded in the “Virtual Column®” software, using the so called "porting" methodology [5]. The Virtual Column software allows the simulation of IC separations performed under a wide range of experimental conditions (*e.g.*, analytes of interest, eluent type, column type, temperature, flow-rate) to identify the optimal eluent and column conditions for a desired separation. This simulation is based on the application of mathematical retention models applied to an extensive database of experimentally-determined analyte retention data embedded in the software [6]. This database covers over 150 anions, cations, and carbohydrate species as analytes, 20 columns, 5 eluent types, 2 column diameters, and 3 temperatures, comprising in total 23,040 datapoints. It is the wide scope of this database which makes Virtual Column such a powerful tool for the development of IC separation

methods, but in turn it is also the validity of these retention data which determines the accuracy of the simulated chromatograms. This retention database was constructed around 10 years ago, so the simulation and optimisation for IC separations can cause errors when the predicted separations are applied on recently produced columns. These errors can result from changes in column behaviour due to either batch-to-batch variability in the production process or the manufacture of new column versions. Errors can also result when the predicted separations are applied to used columns which may have a different ion-exchange capacity to the column on which the original retention data were acquired. With this in mind, a porting methodology to update the retention databases was developed to improve the accuracy of retention prediction by reflecting column-related changes, such as ion-exchange selectivity coefficients and ion-exchange capacity [5].

According to the previously developed porting methodology [5], the retention data embedded into Virtual Column are recalibrated for the entire set of analytes on each particular column by using porting equations which relate existing (or embedded) data to new retention data. In this porting method, new retention data were obtained experimentally by conducting isocratic separations using two representative ions (chloride and thiosulfate) on the desired column under three eluent concentrations. Porting equations describing the changes in retention data for these two ions are then derived and are applied to all ions in the database. The general principle of this approach is that any changes in retention observed for chloride and thiosulfate can be generalized across all ions in the database. Although the porting procedure generally improved the retention prediction accuracy on new columns, such as AS20 and AS11HC columns, the retention prediction accuracy for some columns, such as AS19 column, was poorer than expected. This is attributed to some deficiencies in the porting procedure.

In this study, the accuracy of the porting procedure has been improved by increasing the number of marker anions used to derive the porting equations from two to six. Subsequently, the modified porting method (MPM) has been validated using three newly manufactured Thermo Fisher Scientific columns (AS20, AS19, and AS11-HC). For the validation, the values of mean absolute percentage errors (MAPEs) in the prediction of the retention times were compared in terms of the data types (original embedded data, ported data using the current porting method (CPM), and ported data using the MPM) employed by the mathematical retention model. The accuracy of the retention prediction was then illustrated by plots which show predicted versus measured retention times. Finally, isocratic separations for 13 ions were performed on an AS20 column which had been used for around more than 1500 runs, under three different eluent concentrations. The MPM resulted in a more precise and robust recalibration technique for the update of the retention databases on a wide range of columns compared to the CPM.

3.2. Materials and Methods

3.2.1 General

The isocratic retention data used in this work, which are embedded in the Virtual Column software, had been acquired previously as outlined in reference [7]. These isocratic data were collected at different times using different instruments and columns from different manufacturing batches. Therefore, any comparisons of data made between the isocratic measurements will include variability between instruments and column batches.

3.2.2 Instrumentation

All analyses were carried out using a Dionex (Sunnyvale, CA, USA) ICS-3000 Ion Chromatography system consisting of a dual gradient pump unit (Dionex ICS-3000 DP),

a dual eluent generator unit (Dionex ICS-3000 EG), a dual suppressed conductivity detector compartment (Dionex ICS-3000 DC) and an autosampler (Dionex AS). Separation was performed on Dionex IonPac AS20, AS19, and AS11HC columns (all 250 mm × 4 mm i.d.) with their associated guard columns (all 50 mm × 4 mm i.d.) at a column temperature of 30°C. A Dionex EluGen® cartridge (EGC II KOH) followed by a Dionex CR-ATC ion trap column were employed to generate electrolytically each eluent composition and a Dionex ASRS 300 4 mm suppressor was used for eluent suppression. The analytes were detected by suppressed conductivity at 35°C. An injection volume of 10 μL and an eluent flow-rate of 1.0 mL/min were used throughout this work. Instrument control and data acquisition were performed using Chromeleon® chromatography management software (version 6.80). The following eluent compositions were used to collect isocratic data for the porting and its validation on desired column: 20, 35, and 65 mM hydroxide eluents on AS20 column; 15, 25 and 40 mM on AS19 column; 16, 30 and 45 mM on AS11HC. All experimental points were carried out in triplicate, of which the averaged values were used as the experimental data. The mean and maximum relative standard deviations (RSD) for experimental points deriving the porting equations on AS20, AS19 and AS11HC columns were 0.05% and 0.2%, 0.1% and 0.3% and 01% and 0.8%, respectively.

3.2.3 Void time measurement

The column void time t_0 for the derivation of the porting equations in this work was carefully obtained from the minimum in the water dip peak by using the following equation:

$$t_0(\text{column}) = t_0(\text{analytical column} + \text{guard column} + \text{tubing existing in the IC system}) - t_0(\text{tubing existing in the IC system}) \quad \textbf{(3.1)}$$

The extra column void, t_0 (tubing existing in the IC system) in **Eq. 3.1**, was added back when we calculate the predicted retention times for analytes.

3.3. Results and Discussion

3.3.1 Column void time

Since the equations used to predict retention in Virtual Column are based on retention factors, accurate measurement of the column void time was essential. Here, the void time with the columns connected to the IC system was measured using the water dip peak in the chromatograms of marker anions which had been selected to derive the porting equations. The void time without any column connected to the IC system was then obtained by injecting water as a sample with the suppressor turned off, where the water peak was observed as a dip in the baseline. Careful measurements of the column void time showed that the observed value varied slightly as the eluent concentration was altered. For example, the void times at the eluent concentrations of 20, 35, and 65 mM on the AS20 column were 2.43, 2.45 and 2.48 min, respectively. This is possibly due to the effect of osmotic pressure on the polymer-based ion-exchange stationary phases. For example, when higher eluent concentrations are used, pores within the column can be swollen by water inflow due to osmotic pressure. This finding is also supported by several examples which have recently been reported [8, 9]. For example, it was reported that the pore volume (which is determined by subtracting the interstitial volume from the void volume) in a polymer-based reversed-phase column was changed with mobile phase composition [8] and the pore volume was influenced by eluent composition, temperature, or sample type [9]. Additionally, we have observed that the porting accuracy was improved when the void time adjusted by eluent concentration was applied to the calculation, compared to the use of a constant void time regardless of eluent concentration. As a result, void time measurement was performed separately for each eluent concentration.

3.3.2 Modified porting procedure

Simulation of IC separations under isocratic eluent conditions using Virtual Column is performed based on both the linear solvent strength (LSS) model and two prediction coefficients (*a* and *b*) derived from the LSS model [7].

$$\log k = a - b \log[E^{y-}] \quad \textbf{(3.2)}$$

where *a* and *b* are the intercept and the slope of the model equation, respectively [7, 10, 11]. The *a* and *b* values are determined experimentally for each analyte by fitting the logarithm of retention factor (*k*) to the logarithm of the eluent concentration for three isocratic mobile phase conditions. These *a* and *b* values are then used to predict retention times for all eluent compositions under isocratic, gradient elution, or multi-step elution conditions comprising sequential isocratic and linear gradient steps, based on the corresponding embedded mathematical retention models [7]. Virtual Column 2 provides good agreement between predicted and measured retention times, resulting in reliable optimization and simulation of IC separations. The mean absolute percentage errors (MAPEs) on AS9-HC, AS14A, and AS4A-SC column were 2, 3, and 1%, respectively and in most cases were within 5% [6].

The CPM used to update *a* and *b* values embedded in Virtual Column is described briefly below. Please refer to reference [5] for a more thorough explanation. Retention times for two representative ions (chloride and thiosulfate) were measured experimentally under three eluent concentrations on the column which is the subject of the porting. Subsequently, *a* and *b* values for these two ions were derived from the LSS model and compared with the embedded *a* and *b* values for the same ions. As a result, porting equations, which express the relationships between the embedded and new data regarding each *a* and *b* value, were determined. Finally, *a* and *b* values for the entire set of analytes for that column in the embedded database were updated using these porting equations.

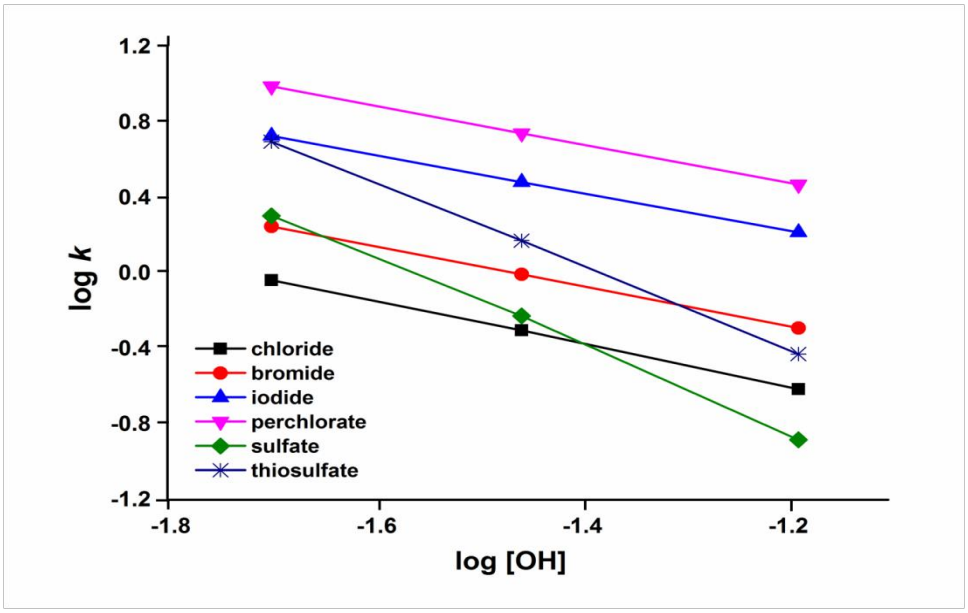
Based on the CPM [5], the porting procedure was modified by employing six anions (chloride, bromide, iodide, perchlorate, sulfate and thiosulfate) as marker ions to derive the porting equations. These six marker ions were selected to include four monovalent ions (chloride, bromide, iodide, and perchlorate) and two divalent ions (sulfate and thiosulfate), among which three are polarizable ions (iodide, perchlorate, and thiosulfate). Essentially, by employing more ions and including a range of ions with different characteristics, the porting equations can more widely and precisely reflect the changes in column behaviour, resulting in an improvement of the retention prediction accuracy of simulated chromatograms. As an example, **Fig. 3.1** shows how the porting equations were derived on a Thermo Fisher Scientific IonPac AS20 column (250 mm × 4 mm i.d.). **Fig. 3.1A** is a plot of log*k* versus log[OH⁻] for the six representative anions. Experimental *a* and *b* values were obtained from the intercepts and the slopes, respectively. **Table 3.1** lists these experimental data, along with their correlation coefficients (R^2 values). Good linearity was observed for all the marker ions (R^2 >0.999). Subsequently, the experimental values were plotted against the embedded values **(Fig. 3.1B)**, from which the porting equations for *a* and *b* values on the AS20 column were derived:

$$a_{\text{ported}} = 0.983\, a_{\text{embedded}} - 0.025 \quad (R^2 = 0.9999) \qquad \textbf{(3.3)}$$

$$b_{\text{ported}} = 0.988\, b_{\text{embedded}} + 0.019 \quad (R^2 = 1) \qquad \textbf{(3.4)}$$

High values of correlation coefficients in the porting equations for the *a* and *b* values were obtained as shown in the **Eq. 3.3** and **3.4**, which supports the porting of the original databases. Similarly, the porting equations for AS19 and AS11HC were derived and are listed in **Table 3.2**. Only five anions (perchlorate was deleted) were employed to derive the porting equation on the AS11HC column, as there are no embedded data for this particular ion on this column.

A)



B)

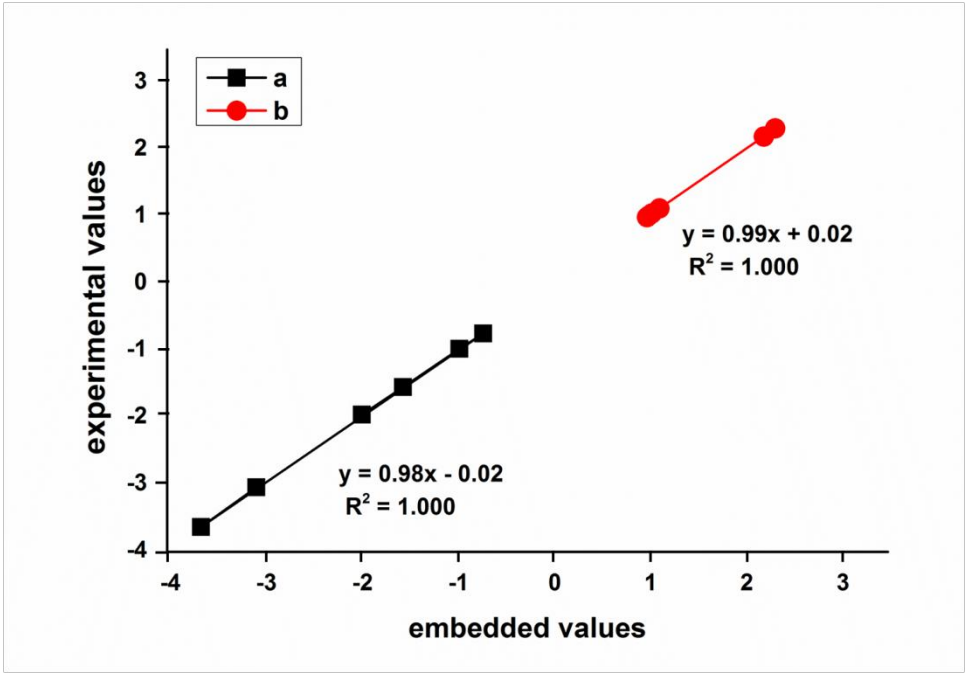


Figure 3.1 Modified porting procedure. (A) Plot of log*k* vs. log[OH^-] for the marker ions is used to determine experimental *a* (intercept) and *b* (slope) for deriving the porting equations. Conditions: column: 4 mm × 250 mm AS20 with AG20 guard column; eluent: 20, 35, and 65 mM KOH; flow rate: 1.0 mL/min. Associated slopes, intercepts and R^2 values are shown in Table 3.1. (B) Experimental values *vs.* embedded values for the six marker ions to derive the porting equations.

Table 3.1 Experimental *a* and *b* values for six anions used to derive porting equations shown in Fig. 3-1A.

Compound	*a*	*b*	R^2
Chloride	- 1.952	1.129	0.9997
Bromide	- 1.540	1.053	1.0000
Iodide	- 0.969	0.999	1.0000
Perchlorate	- 0.737	1.017	1.0000
Sulfate	- 3.627	2.321	0.9990
Thiosulfate	- 3.035	2.199	0.9999

For conditions see Fig. 3-1.

Table 3.2 Summary of porting equations on the AS20, AS19 and AS11HC columns.

Column	Porting equation (a)	Porting equation (b)
AS20	$a_{ported} = 0.983\ a_{embedded} - 0.025$	$b_{ported} = 0.988\ b_{embedded} + 0.019$
AS19	$a_{ported} = 1.02\ a_{embedded} + 0.055$	$b_{ported} = 0.994\ b_{embedded} + 0.007$
AS11HC	$a_{ported} = 0.991\ a_{embedded} + 0.002$	$b_{ported} = 0.994\ b_{embedded} - 0.02$

3.3.3 Validation of the ported database

The MPM was validated by investigating the MAPEs of the predictions of the retention times. The MAPEs were determined by averaging the absolute values of the percentage errors (APE) of retention times, which were calculated by the following equation:

$$\text{APE} = \left| 100 \times \frac{(t_{\text{r\_predicted}} - t_{\text{r\_experimental}})}{t_{\text{r\_experimental}}} \right| \quad \textbf{(3.5)}$$

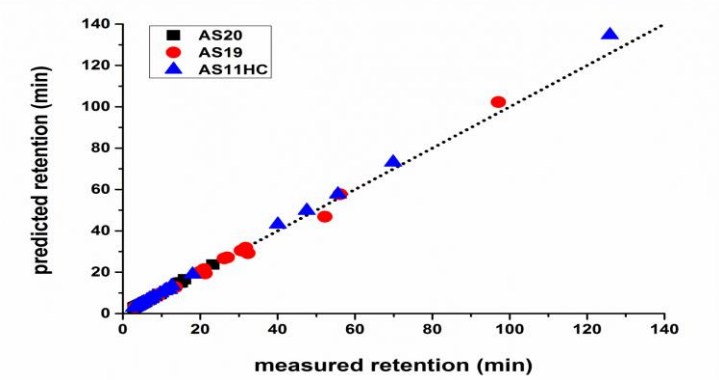
where the predicted retention times were calculated by employing either the ported *a* and *b* values or the embedded *a* and *b* values in the LSS model. To investigate whether the MPM can improve the accuracy of retention prediction, the MAPEs of the retention times obtained using ported data were compared with those using the embedded data. The following ten ions were used as validation test analytes: fluoride, nitrite, nitrate, thiocyanate, carbonate, malonate, oxalate, succinate, molybdate, and phosphate. Retention times for these test anions were measured under three eluent concentrations and subsequently compared with the retention times predicted using both embedded *a* and *b* values and the *a* and *b* values obtained using the CPM and the MPM. **Table 3.3** compares the MAPE values in terms of columns and data types used during the calculation. The data show that the MPM provided the best prediction accuracy, with MAPEs of <1.3% for all three columns. **Fig. 3.2** shows the correlation between measured and predicted retention times for the ten test ions, measured under three different eluent concentrations. The superiority of the MPM is clearly evident, with **Fig. 3.2C** exhibiting the lowest degree of scatter of the points and the best fit to the 45 degree line (dotted line in **Figs. 3.2A-C**).

To check the robustness as well as the prediction accuracy of the MPM, an AS20 column which had been used for more than 1500 runs was selected for the isocratic separation of 13 ions. The test analytes (acetate, benzoate, bromate, carbonate, clorate,

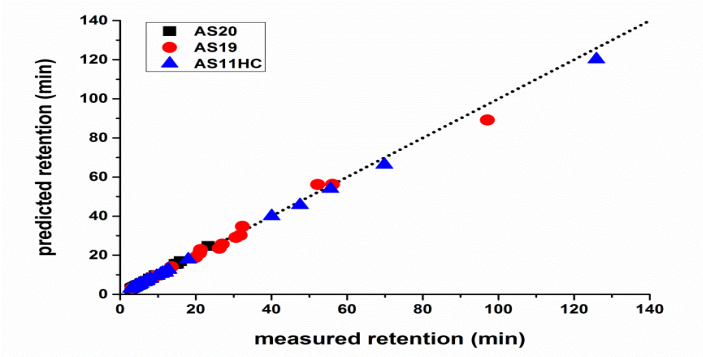
Table 3.3 Mean absolute percentage errors (MAPE) between experimental and predicted retention times for 10 test anions.

Column	Embedded data	Ported data (CPM)	Ported data (MPM)
AS20	2.0	1.8	1.3
AS19	3.2	3.7	1.3
AS11HC	2.6	1.5	1.3

A)



B)



C)

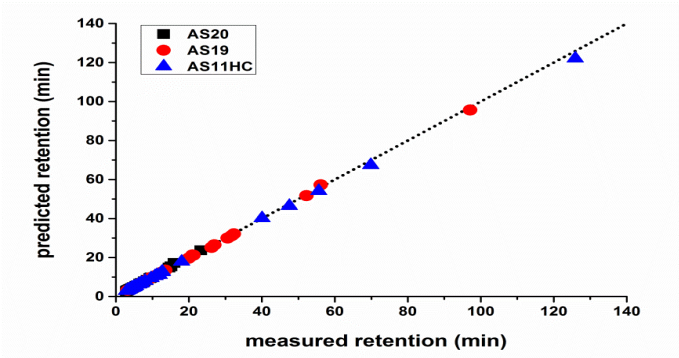


Figure 3.2 Predicted retention times versus measured retention times of the 10 test anions under 3 different eluent conditions on 3 different columns. The embedded *a* and *b* values (A), the ported values using the current porting method (B), and the ported values using the modified porting method (C) were used to calculate the predicted retention times.

chromate, fluoride, formate, nitrate, nitrite, phosphate, sulfate, and thiocyanate) are inorganic and small organic anions related to the identification of explosives [12]. They were separated on the AS20 column using eluent concentrations of 25, 40 and 50 mM OH^-, which were different from the concentrations used to derive the porting equations. It is noteworthy that the MAPE using MPM has an acceptable level of accuracy *i.e.*, 3%, although the column behaviour had been altered substantially due to its numerous uses. The changes of the column behaviour resulted in a large increase of prediction error of 10% when the embedded data were used for modelling and 7% when the CPM was used. **Fig. 3.3** shows the errors for each individual ion in the test mixture (listed in elution order) and this figure demonstrates that for the embedded data and the data obtained by the CPM, errors are greatest for the longer retained species. This can be attributed to loss of ion-exchange capacity of the column during use. By contrast, the errors for the MPM were relatively constant for each of the test anions, with the exception of thiocyanate.

3.4. Conclusions

In this study an existing porting methodology for updating an extensive retention database embedded in the Virtual Column software has been modified and improved. Essentially, with the modified porting procedure, good agreement between measured and predicted retention times was observed, with the MAPE <1.3% on all three columns tested (AS20, AS19 and AS11HC). It is concluded that the modified porting approach can be applied to a wide range of columns to update IC retention data, giving better retention prediction accuracy compared to the previous porting method. Thus, updating retention prediction data via the modified porting methodology can provide more reliable simulations and *in silico* optimization of IC separations.

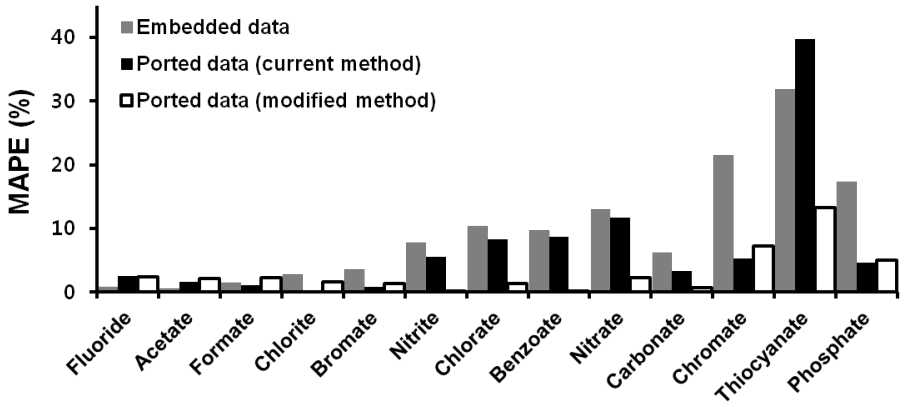


Figure 3.3 Mean absolute percentage errors (MAPEs) for the 13 ions under 3 different isocratic eluent conditions using 25, 40 and 50 mM KOH as eluent. The MAPEs for total 13 ions using embedded data, ported data using current porting method, and ported data using the modified porting method were 10, 7 and 3%, respectively.

3.5. References

[1] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey, Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part I: Isocratic separation, Eur. J. Pharm. Biopharm. 66 (2007) 475-482.

[2] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey, Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments, Eur. J. Pharm. Biopharm. 68 (2008) 430-440.

[3] L.M. Blumberg, M.S. Klee, Method translation and retention time locking in partition GC, Anal. Chem. 70 (1998) 3828-3839.

[4] M.S. Klee, L.M. Blumberg, Theoretical and practical aspects of fast gas chromatography and method translation, J. Chromatogr. Sci. 40 (2002) 234-247.

[5] B.K. Ng, R.A. Shellie, G.W. Dicinoski, C. Bloomfield, Y. Liu, C.A. Pohl, P.R. Haddad, Methodology for porting retention prediction data from old to new columns and from conventional-scale to miniaturised ion chromatography systems, J. Chromatogr. A 1218 (2011) 5512-5519.

[6] J.E. Madden, M.J. Shaw, G.W. Dicinoski, N. Avdalovic, P.R. Haddad, Simulation and optimization of retention in ion chromatography using virtual column 2 software, Anal. Chem. 74 (2002) 6023-6030.

[7] R.A. Shellie, B.K. Ng, G.W. Dicinoski, S.D.H. Poynter, J.W. O'Reilly, C.A. Pohl, P.R. Haddad, Prediction of analyte retention for ion chromatography separations performed using elution profiles comprising multiple isocratic and gradient steps, Anal. Chem. 80 (2008) 2474-2482.

[8] B. Trathnigg, M. Veronik, A. Gorbunov, Looking inside the pores of a chromatographic column - I. Variation of the pore volume with mobile phase composition, J. Chromatogr. A 1104 (2006) 238-244.

[9] M. Wang, J. Mallette, J.F. Parcher, Comparison of void volume, mobile phase volume and accessible volume determined from retention data for oligomers in reversed-phase liquid chromatographic systems, J. Chromatogr. A 1218 (2011) 2995-3001.

[10] P.R. Haddad, P.E. Jackson, Ion chromatography: principles and applications, in: Journal of Chromatography Library, vol. 46, Elsevier, Amsterdam, The Netherlands, 1990, p. 135.

[11] J.E. Madden, N. Avdalovic, P.E. Jackson, P.R. Haddad, Critical comparison of retention models for optimisation of the separation of anions in ion chromatography III. Anion chromatography using hydroxide eluents on a Dionex AS11 stationary phase, J. Chromatogr. A 837 (1999) 65-74.

[12] C. Johns, R.A. Shellie, O.G. Potter, J.W. O'Reilly, J.P. Hutchinson, R.M. Guijt, M.C. Breadmore, E.F. Hilder, G.W. Dicinoski, P.R. Haddad, Identification of homemade inorganic explosives by ion chromatographic analysis of post-blast residues, J. Chromatogr. A 1182 (2008) 205-214.

Chapter 4

Quantitative structure-retention relationships applied to the linear solvent strength model

4.1. Introduction

Quantitative structure-retention relationships (QSRRs) can provide a powerful alternative to the conventional trial-and-error approach used in chromatographic method development and can lead to substantial savings in time and cost. One common QSRR approach involves the generation of statistically-derived mathematical relationships between molecular descriptors of analytes calculated from molecular modelling of chemical structures and chromatographic retention parameters, and the subsequent use of these relationships to predict the retention times of new analytes not included in the model generation [1]. Retention time predictions of new analytes performed in this way can simplify the selection of broad chromatographic conditions, based only on the experimentation needed to compile the retention database used to develop the QSRRs [2]. In particular, this type of modelling can play an important role in rapid screening of a large number of potential columns for a desired separation, *i.e.*, for the "scoping" stage of method development. Relatively higher errors in the prediction of retention times can be acceptable at the scoping stage, compared to the later detailed optimisation stage where precise optimal mobile phase conditions need to be identified.

A number of studies regarding the prediction of retention times using QSRRs have been reported in the area of liquid chromatography (LC), such as in reversed-phase LC (RPLC) [3-6], hydrophilic interaction LC (HILIC) [5], and ion-exchange chromatography (IEC) [7-12], as well as gas chromatography [13-16]. The QSRR models derived in these studies have mainly applied multiple linear regression (MLR) [3-6, 10-11, 13-16],

although other modelling methods, such as partial least squares (PLS) [6, 7, 11, 13, 15, 16] and support vector machine (SVM) [8] regression, have also been used. Additionally, these QSRRs require selection techniques to reduce the number of variables (*i.e.*, molecular descriptors) used since more than 4000 theoretical molecular descriptors are available to build the QSRR models, including a large number of descriptors which are redundant [17-21]. Evolutionary algorithms (EA), such as genetic algorithms, have been widely employed as a variable selection technique [3, 12, 21], as well as alternative methods such as Monte Carlo uninformative variable elimination (MC-UVE), iteratively retaining informative variables (IRIV), competitive adaptive reweighted sampling (CARS), and variable iterative space shrinkage approach (VISSA) [20].

QSRR has been applied in IEC for retention time predictions of proteins [7-9], ionic liquids [10] and arylakyl amines [22, 23]. Additionally, QSRR models to predict the gradient retention of carbohydrates in IEC have been derived by various linear regression methods (stepwise MLR, PLS, and uninformative variable elimination-partial least squares (UVE-PLS)) [11], as well as artificial neural networks (ANN) [12]. Another study regarding the QSRR prediction of acidic analytes in strong anion-exchange chromatography has been reported, where ANN modelling using pre-selected physicochemical properties as descriptors was employed to elucidate various retention mechanisms [24].

When IEC is applied to the separation of inorganic and small organic ions, the technique is generally referred to as ion chromatography (IC). In IC, the linear solvent strength (LSS) model is usually employed to relate isocratic retention to the eluent concentration as follows [25, 26]:

$$\log k = a - b \log[E^{y-}] \qquad \textbf{(4.1)}$$

where k is the retention factor, $[E^{y-}]$ is the molar concentration of the eluent competing ion (mol/L), and a and b values are respectively the intercept and the slope. Based on the LSS model, IC separations for inorganic and small organic ions can be simulated and optimized by selecting the correct eluent concentration. The a and b values for a given ion on a specific column are firstly estimated by fitting retention data (k), measured under three eluent concentrations, to the LSS model. Once an estimate of the a and b values is obtained, the LSS model can then be used to predict the retention times of the analytes in the database for all eluent compositions. In addition to isocratic retention prediction, gradient retention, as well as multi-step elution profiles comprising sequential isocratic and gradient steps can also be predicted accurately using the a and b values in the LSS model [26]. This procedure can only provide optimal eluent concentrations for the ions where the a and b values are already in the database. Thus, a QSRR approach wherein the a and b values in **Eq. 4.1** are predicted for any new analytes not included in the database, based only on their chemical structures, is particularly attractive since this would enable prediction of their retention times for all isocratic, gradient and multi-step eluent compositions. To perform such QSRR modelling, databases of a and b values for a range of analytes and stationary phases are required. **Figure 4.1** outlines the general procedure for the QSRR modelling developed in this study. Firstly, datasets are created, composed of a and b values obtained from retention data over a variety of analytes, columns, and eluent concentrations. Secondly, molecular descriptors containing structural information for the analytes are calculated for all the analytes in the datasets using molecular modelling software such as Dragon. Thirdly, the dataset is split into two groups - the training and external test sets. Fourthly, an evolutionary algorithm (EA) using analytes in the training set is utilised to select the most significant descriptors related to the IC retention. Finally, QSRR models for a and b values are built using the MLR. To validate the developed models, external

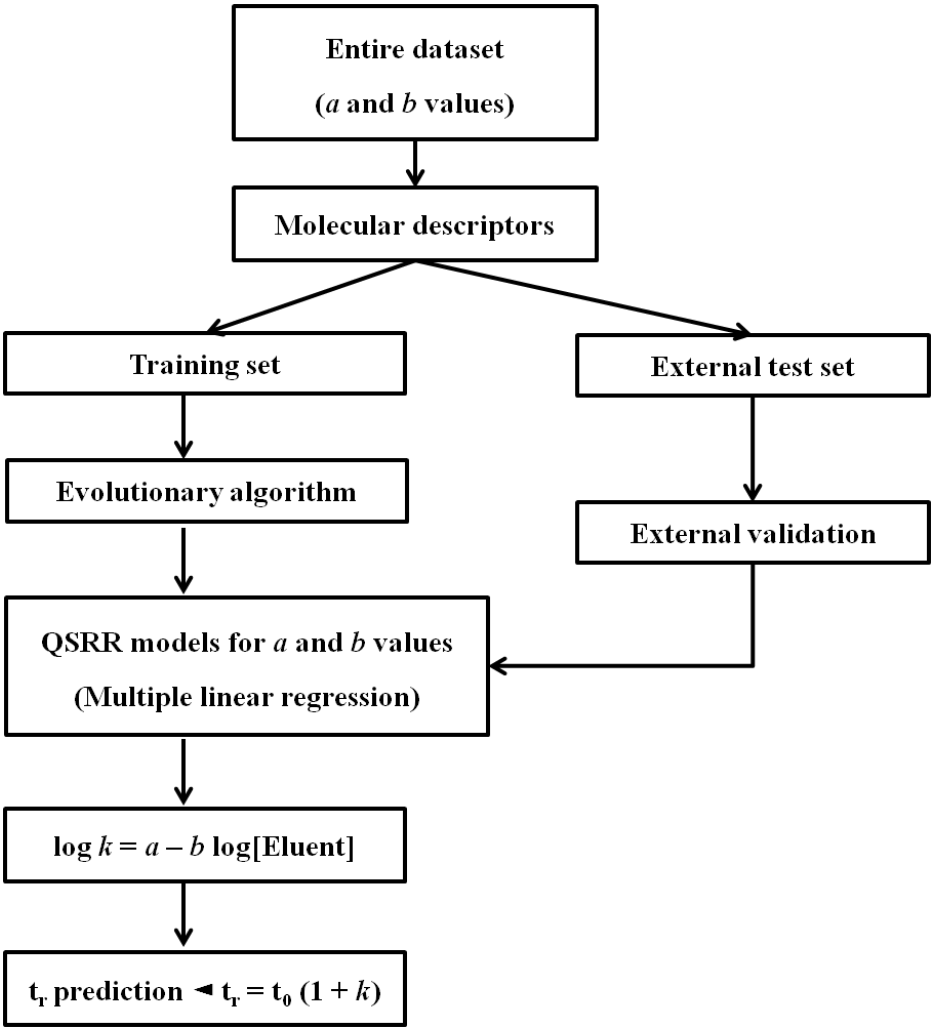


Figure 4.1 Schematic overview of QSRR modelling by MLR

validation is performed using the external test set. Once successful QSRR models for the *a* and *b* values are created, the retention times of target anions can be predicted by substituting predicted *a* and *b* values (from the molecular structure of the unknown anion) into the LSS model.

In this chapter, QSRR models for the *a* and *b* values in the LSS model have been successfully created by EA-MLR for inorganic and small organic ions on three Thermo Fisher Scientific columns (49, 41 and 40 compounds on AS20, AS19 and AS11HC columns, respectively). The optimal number of relevant descriptors for the QSRR models was determined by considering the values of coefficient of determination (R^2) and root-mean-square error (*RMSE*) from the models observed when the number of descriptors was increased progressively. Statistically significant and accurate QSRR models were then generated by the MLR method, correlating retention parameters (*a* and *b* values) to the optimal subset of molecular descriptors. Finally, external validation showed that the constructed QSRR models could be applied as predictive tools in IC. Furthermore, predicted retention times of the anions, calculated by inserting the predicted *a* and *b* values into the LSS model **[Eqn. (4.1)]**, were compared with their measured retention times and showed good agreement. The results obtained in this chapter show that QSRR models for *a* and *b* values can properly predict the retention times of inorganic and small organic anions in IC.

4.2. Materials and Methods

4.2.1. Datasets

Retention data of inorganic and small organic anions used in this chapter were obtained by recalibrating the retention data embedded in the Thermo Fisher Virtual Column™ software utilising the modified "porting" methodology described in chapter 3 [27]. Using this porting strategy the retention data were updated to address possible

changes in retention on newly produced columns due to batch-to-batch variability. Data sets on three columns (4 mm I.D. Dionex IonPac AS20, AS19 and AS11HC) at a column temperature of 30°C were selected **(Table 4.1)** from the extensive retention databases embedded in the Virtual Column software [26]. The *a* and *b* values were derived from the LSS model **(Eq. 4.1)** using retention time (t_R) data collected under three isocratic eluent concentrations at 30°C. The datasets were split into training and external test sets in 9:1 ratios. The training sets were used to build QSRR models and the external test sets used to evaluate the predictive ability of all generated models.

4.2.2. Molecular descriptors

The molecular structures of the analytes were drawn using MarvinSketch version 6.2.1 (ChemAxon, Budapest, Hungary) and were then entered into the Spartan '14 (version 1.1.4) software (Wavefunction Inc., CA, USA). Geometry optimisations of the molecular structures were performed using AM1 semi-empirical methods, following the generation of initial 3D coordinates by the Merck Molecular Force Field (MMFF). These optimized 3D structures were then imported into the Dragon 6.0 software (Talete, Milano, Italy) to calculate the molecular descriptors. The original set of 4885 descriptors was reduced to a subset of 182 descriptors using automatic screening by applying the following criteria. First, the descriptors with constant (or nearly-constant) values, as well as those for which some analytes were missing values were excluded. Second, descriptors with a standard deviation less than 0.001 were also removed. Finally, where two descriptors showed an (absolute) pairwise correlation $\geq$0.7, one was eliminated. The 182 descriptors generated by Dragon were entered into the EA for the selection of the most important variables (*i.e.*, feature selection).

4.2.3. Feature selection using evolutionary algorithm

The feature selection to extract the most significant descriptors that generate the best-fitted models was implemented with *in-house* developed EA software, broadly

Table 4.1 Compound names and their measured *a* and *b* values on AS20, AS19 and AS11HC columns at 30°C.

AS20							
No.	**Compounds**	***a***	***b***	**No.**	**Compounds**	***a***	***b***
1	Acetate	-2.63	1.32	**26**	Methanesulfonate	-2.30	1.19
2	Acrylate	-2.33	1.22	**27**	n-Butyrate	-2.43	1.22
3	Benzenesulfonate	-1.37	1.01	**28**	Nitrate	-1.48	1.06
4	Benzoate	-1.47	1.02	**29**	Nitrite	-1.73	1.08
5	Bromoacetate	-2.05	1.13	**30**	n-Valerate	-2.27	1.17
6	Butanesulfonate*	-2.10	1.14	**31**	Octanesulfonate	-1.22	0.95
7	Carbonate	-3.59	2.25	**32**	Oxalate	-3.35	2.24
8	Chlorate	-1.63	1.07	**33**	p-Chlorobenzenesulfonate	-0.98	0.98
9	Chloroacetate	-2.16	1.17	**34**	Pentanesulfonate	-1.95	1.11
10	Citrate	-4.52	3.37	**35**	Perchlorate	-0.72	1.02
11	Dibromoacetate	-1.57	1.04	**36**	Phthalate	-3.04	2.17
12	Dichloroacetate	-1.71	1.06	**37**	Propanesulfonate	-2.25	1.19

13	Difluoroacetate	-2.07	1.15	**38**	Propionate*	-2.54	1.27
14	Ethanesulfonate	-2.37	1.23	**39**	Pyruvate	-2.35	1.23
15	Fluoroacetate*	-2.46	1.27	**40**	Quinate	-3.10	1.53
16	Formate	-2.43	1.27	**41**	Sorbate	-1.73	1.04
17	Fumarate	-2.72	2.10	**42**	Succinate	-3.24	2.18
18	Glutarate*	-3.22	2.16	**43**	Sulfate	-3.61	2.32
19	Glycolate	-2.88	1.45	**44**	Sulfite	-3.56	2.27
20	Heptanesulfonate	-1.49	1.00	**45**	Tartrate	-3.37	2.24
21	Hexanesulfonate	-1.71	1.04	**46**	Thiosulfate	-3.05	2.20
22	Lactate	-2.81	1.39	**47**	Tribromoacetate	-0.96	0.98
23	Malate	-3.34	2.21	**48**	Trichloroacetate	-1.21	0.99
24	Malonate*	-3.42	2.24	**49**	Trifluoroacetate	-1.70	1.06
25	Methacrylate	-2.23	1.17				

AS19

No.	**Compounds**	***a***	***b***	**No.**	**Compounds**	***a***	***b***
1	Acetate	-2.09	1.11	**22**	Methanesulfonate	-1.94	1.10

2	Acrylate	-1.92	1.09	**23**	n-Butyrate*	-2.03	1.09
3	Benzoate	-1.28	1.02	**24**	Nitrate	-1.24	1.05
4	Bromoacetate	-1.74	1.07	**25**	Nitrite	-1.48	1.05
5	Butanesulfonate*	-1.78	1.07	**26**	n-Valerate	-1.93	1.08
6	Carbonate	-2.98	2.11	**27**	Oxalate	-2.79	2.13
7	Chlorate	-1.38	1.05	**28**	Pentanesulfonate	-1.65	1.05
8	Chloroacetate	-1.81	1.07	**29**	Perchlorate	-0.47	1.04
9	Dibromoacetate	-1.36	1.03	**30**	Phthalate	-2.61	2.13
10	Dichloroacetate	-1.48	1.04	**31**	Propanesulfonate	-1.87	1.08
11	Difluoroacetate	-1.75	1.07	**32**	Propionate*	-2.08	1.11
12	Ethanesulfonate	-1.94	1.09	**33**	Pyruvate	-1.93	1.09
13	Formate	-1.94	1.09	**34**	Quinate	-2.27	1.15
14	Fumarate	-2.37	2.10	**35**	Sorbate	-1.53	1.03
15	Glutarate*	-2.75	2.09	**36**	Succinate	-2.75	2.10
16	Glycolate	-2.15	1.13	**37**	Tartrate	-2.83	2.13
17	Hexanesulfonate	-1.48	1.03	**38**	Thiosulfate	-2.58	2.16

18	Lactate	-2.17	1.13	**39**	Tribromoacetate	-0.76	1.01
19	Malate	-2.82	2.11	**40**	Trichloroacetate	-1.02	1.02
20	Malonate	-2.86	2.12	**41**	Trifluoroacetate	-1.46	1.04
21	Methacrylate	-1.88	1.08				

AS11HC

No.	**Compounds**	***a***	***b***	**No.**	**Compounds**	***a***	***b***
1	Acetate	-2.73	1.18	**21**	Methacrylate	-2.00	1.02
2	Acrylate	-2.23	1.05	**22**	Methanesulfonate	-2.30	1.09
3	Benzoate	-1.00	0.96	**23**	n-Butyrate*	-2.29	1.05
4	Bromoacetate	-1.73	1.00	**24**	Nitrate	-1.21	1.00
5	Butanesulfonate*	-1.61	0.99	**25**	Nitrite	-1.59	1.01
6	Carbonate	-3.33	2.12	**26**	n-Valerate	-1.94	1.00
7	Chlorate	-1.16	1.00	**27**	Oxalate	-3.05	2.13
8	Chloroacetate	-1.94	1.04	**28**	Pentanesulfonate	-1.21	0.97
9	Dibromoacetate	-0.91	0.97	**29**	Phthalate	-2.59	2.12
10	Dichloroacetate	-1.20	0.98	**30**	Propanesulfonate	-1.95	1.03

11	Difluoroacetate	-1.87	1.02	**31**	Propionate*	-2.51	1.10
12	Ethanesulfonate	-2.22	1.07	**32**	Pyruvate	-2.23	1.07
13	Formate	-2.43	1.10	**33**	Sorbate	-1.27	0.97
14	Fumarate	-2.95	2.08	**34**	Succinate	-3.36	2.11
15	Glutarate*	-3.40	2.10	**35**	Sulfate	-3.21	2.16
16	Glycolate	-2.78	1.20	**36**	Sulfite	-3.24	2.14
17	Hexanesulfonate	-0.81	0.95	**37**	Tartrate	-3.34	2.14
18	Lactate	-2.85	1.22	**38**	Thiosulfate	-2.50	2.15
19	Malate	-3.38	2.12	**39**	Trichloroacetate	-0.37	0.96
20	Malonate	-3.30	2.13	**40**	Trifluoroacetate	-1.30	0.98

* Compounds used in the external test set.

based on the procedure described in [28]. The EA is a stochastic optimisation technique mimicking the evolution theory. The optimisation is based on the evolution of models where a random starting population of models reaches an optimal solution after a number of generations through cross-over, mutation and selection. As a result, the significant set of descriptors, which are those that optimise the performance of a model response, is chosen to build QSRR models [3]. The EA parameters applied in this work were as follows. The size of the population was 50, the probability of cross-over and mutation were 60% and 2%, respectively, and the number of generations when the algorithm stopped was limited to 200 [28].

4.2.4. QSRR modelling by EA-MLR

QSRR models for both *a* and *b* values in the LSS model were built by MLR in MATLAB R2015a (MathWorks, Natick, MA, USA) software. The training sets, used to generate the QSRR models, represented 90% of the compounds in each dataset (44, 37 and 36 anions, on the AS20, AS19 and AS11HC columns, respectively). External validation was performed on the remaining 10% of the data set (*i.e.*, the external test set). Both Q^2- and root mean square error of prediction (*RMSEP*)-values were calculated to evaluate the predictive power of the developed QSRR models.

4.3. Results and Discussion

4.3.1. Determination of the optimal number of molecular descriptors

The selection of reliable descriptors is an essential step in the generation of accurate QSRR models. In constructing QSRR models for retention prediction, molecular descriptors are typically obtained either from well-known physicochemical properties or selected from a large pool of theoretical descriptors [19]. This study followed the latter approach. To determine the optimal number of descriptors, the EA-MLR was firstly performed with an increasing number of descriptors [13]. Maximal

squared correlation coefficient (R^2) and minimal root mean squared error (*RMSE*) of predicted *versus* experimental *a* and *b* values, for training and internal test sets were used as optimisation criteria. As a result, the number of significant descriptors resulting in the best-fitted models and the R^2 and *RMSE* values of those models, for both training and internal test sets on the three columns, are given in **Table 4.2**.

Although all highly correlated and zero variance descriptors were already removed using the Dragon software, multi-collinear descriptors can possibly arise in building MLR models. To avoid the over-fitting of the models, the multi-collinearity among the chosen descriptors was further investigated and the small number of existing multi-collinear descriptors was then removed [29]. Multi-collinearity occurs when there are high correlations among two or more independent variables (descriptors) in multiple linear regression models. This can lead to biased regression coefficients in the models, followed by the misinterpretation of the models. As a measure of multi-collinearity, the variance inflation factor (VIF) was used, defined as:

$$\mathrm{VIF} = \frac{1}{1-\mathrm{R}_\mathrm{i}^2} \qquad \textbf{(4.2)}$$

where, R_i^2 is the R^2 for the i^{th} independent variable (descriptor) regressed on the other independent variables (chosen descriptors) in a MLR model [29]. Since a VIF greater than 10 can cause a serious multi-collinearity [30, 31], either the descriptors with the problematic VIF values or the descriptors with a relatively large number of highly-correlated descriptors (correlation coefficient r >0.5) in the correlation matrix were eliminated. Each time only one descriptor was removed and the correlation matrix and VIF values for the remaining descriptors in the MLR model were then checked. This was repeated until the VIF values for all the remaining descriptors in the model were less than 10. Additionally, the descriptor was added back in the model in case the eliminated descriptor had a significant correlation with the retention. In this way, the optimal numbers of descriptors for the QSRR models were selected, without the

Table 4.2 The optimal number of descriptors selected by EA-MLR and the statistical parameters for best-fitted models.

	No. Desc.		R^2 **(training)**		R^2 **(internal test)**		***RMSE*** **(training)**		***RMSEP*** **(internal test)**	
	a **values**	***b*** **values**	***a*** **values**	***b*** **values**	***a*** **values**	***b*** **values**	***a*** **values**	***b*** **values**	***a*** **values**	***b*** **values**
AS20	25	25	0.996	0.998	0.963	0.975	0.037	0.025	0.222	0.132
AS19	20	15	0.995	0.995	0.990	0.989	0.037	0.033	0.169	0.13
AS11HC	20	20	0.999	0.998	0.993	0.963	0.031	0.028	0.172	0.183

presence of significantly multi-collinear descriptors. As an example, the optimal number of descriptors for *a* and *b* values on the AS19 column were determined to be 16 and 15, respectively, and their correlation matrix and VIF values are given in **Table 4.3** and **4.4**. In addition, **Table 4.5** lists the selected descriptors for *a* and *b* values on the three columns (AS20, AS19 and AS11HC).

4.3.2. QSRR modeling by MLR

QSRR models for both *a* and *b* values were generated by MLR for the anions in the training sets on three columns (**Table 4.1**) using the selected descriptors (**Table 4.5**). Simply, the QSRR models for the *a* and *b* values using MLR can be expressed as:

$$a = c_0 + c_1MD_1 + c_2MD_2 + c_3MD_3 + \ldots + c_mMD_m \quad \textbf{(4.3)}$$

$$b = d_0 + d_1md_1 + d_2md_2 + d_3md_3 + \ldots + d_nmd_n \quad \textbf{(4.4)}$$

where c_0 and d_0 are the intercepts, c_1, ..., c_m, and d_1, ..., d_n are the regression coefficients, MD_1, ..., MD_m, and md_1, ..., md_n are the independent variables (selected descriptors), m and n are the number of the selected descriptors, and *a* and *b* are dependent variables [32]. The resulting regression coefficients of the MLR models derived on AS20, AS19 and AS11HC columns are presented in **Tables 4.6, 4.7** and **4.8**, respectively. The absolute value of a regression coefficient represents the magnitude of independent contribution of each descriptor to the dependent variable [32]. Subsequently, *a* and *b* values for each ion were calculated using the given regression coefficients of the MLR models and the descriptors for the ions *via* **Eq. 4.3** and **4.4** and these calculated (or predicted) values were compared with their corresponding measured values. The constructed QSRR models were assessed using the following statistical parameters: squared correlation coefficients R^2, root mean square error (*RMSE*), Fisher ratio value *F* and the significance of the variables in the model *p*. The *RMSE* was determined *via* the following equation:

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{0.5} \quad \textbf{(4.5)}$$

Table 4.3 The correlation matrix and VIF values for the selected descriptors to build QSRR model for *a* values on the AS19 column.

	H0i	SIC0	MATS6p	E2s	nO	pMAD_B(s	Mor21s	AAC	MEcc	nCconj	MATS6v	SaasC	TS2D_01_	Mor26v	Mor27u	MATS3e	VIF
H0i	1																2.88
SIC0	0.05	1															5.35
MATS6p	-0.17	-0.27	1														2.16
E2s	0.24	0.14	-0.04	1													2.17
nO	0.08	-0.53	0.11	-0.01	1												5.52
SpMAD_B(s)	0.43	0.04	-0.15	0.25	0.42	1											3.06
Mor21s	-0.24	-0.02	-0.02	0.15	0.17	-0.12	1										2.24
AAC	0.31	0.58	-0.02	0.08	-0.18	0.34	-0.24	1									3.64
MEcc	-0.09	0.15	0.07	0.05	-0.07	0.29	-0.43	0.31	1								2.85
nCconj	-0.23	-0.24	0.08	-0.15	-0.06	-0.04	0.15	-0.04	0.13	1							2.04
MATS6v	0.03	0.06	0.07	0.06	0.40	0.17	0.04	0.06	-0.02	0.06	1						2.16
SaasC	-0.12	0.10	-0.47	-0.23	-0.19	-0.06	0.00	-0.02	-0.02	-0.16	-0.35	1					1.87
CATS2D_01_AL	0.11	-0.27	-0.06	-0.05	0.14	-0.18	0.31	-0.50	-0.71	-0.11	0.01	0.03	1				2.96
Mor26v	-0.13	0.19	-0.38	0.04	-0.41	-0.07	0.17	0.05	-0.05	0.37	-0.21	0.16	-0.03	1			2.04
Mor27u	-0.03	0.07	-0.10	0.18	0.18	-0.08	0.44	-0.13	-0.08	-0.12	-0.18	0.06	0.08	-0.02	1		2.15
MATS3e	0.05	-0.30	0.14	-0.45	0.05	-0.40	-0.02	-0.35	-0.11	0.08	0.00	-0.05	0.09	-0.20	0.17	1	2.70

Table 4.4 Correlation matrix and VIF values for the selected descriptors to build QSRR model for *b* values on the AS19 column.

	P_VSA_s_3	LDI	GATS6e	Mor09e	nCp	R1v+	H-052	Mor21s	Eig04_EA(dm)	Mor27u	RDF035v	PDI	Mor30s	ChiA_X	MATS3v	VIF
P_VSA_s_3	1															3.13
LDI	-0.47	1														3.68
GATS6e	0.35	-0.27	1													2.62
Mor09e	-0.21	0.03	0.07	1												2.07
nCp	-0.05	-0.27	0.01	0.29	1											2.54
R1v+	-0.02	-0.26	-0.04	0.24	0.11	1										1.38
H-052	-0.16	0.12	0.02	0.44	0.38	0.01	1									2.01
Mor21s	0.13	0.09	-0.19	-0.17	-0.31	-0.24	-0.37	1								5.29
Eig04_EA(dm)	-0.13	0.29	0.33	0.10	-0.41	-0.14	0.15	-0.16	1							2.82
Mor27u	0.40	-0.31	-0.17	-0.13	-0.15	0.00	-0.32	0.44	-0.29	1						3.62
RDF035v	0.33	-0.28	0.35	-0.01	-0.01	0.01	0.05	-0.14	0.22	-0.37	1					3.39
PDI	0.16	0.06	0.49	0.12	0.29	0.02	0.21	-0.04	0.04	-0.32	0.11	1				2.36
Mor30s	0.30	-0.34	0.16	0.18	-0.15	-0.03	-0.16	0.60	0.14	0.25	0.22	0.03	1			5.35
ChiA_X	-0.55	0.38	-0.16	-0.07	-0.16	-0.06	-0.16	0.08	0.17	-0.01	-0.63	-0.11	-0.16	1		2.88
MATS3v	0.05	0.15	0.09	-0.07	-0.29	0.00	0.09	-0.02	0.04	0.03	0.01	-0.12	-0.29	0.01	1	1.65

Table 4.5 Optimal descriptors selected for the 6 QSRR models.

AS20-*a* values				**AS20-*b* values**			
1	Mor15m	12	Mor17m	1	Mor30m	12	CATS2D_01_AL
2	MATS2e	13	P_VSA_s_3	2	Mor21s	13	SpMax_AEA(dm)
3	MATS4m	14	H-047	3	nCconj	14	RDF020e
4	nO	15	E2p	4	Mor17s	15	SpMAD_B(s)
5	CATS2D_01_AL	16	MEcc	5	E1u	16	RARS
6	GATS6p	17	SpMin7_Bh(i)	6	nN	17	MATS2e
7	SpMax_AEA(dm)	18	MATS8p	7	nF	18	E3p
8	nN	19	Mor04m	8	Mor32i	19	E3u
9	Eig09_AEA(bo)	20	Mor10e	9	L3m	20	ATS8m
10	MW	21	Eig07_EA(dm)	10	RDF035s	21	Mor29u
11	E2s	22	Mor25i	11	nO	22	H-051
AS19 - *a* values				**AS19 - *b* values**			
1	H0i	9	MEcc	1	P_VSA_s_3	9	Eig04_EA(dm)
2	SIC0	10	nCconj	2	LDI	10	Mor27u
3	MATS6p	11	MATS6v	3	GATS6e	11	RDF035v

4	E2s	12	SaasC	4	Mor09e	12	PDI
5	nO	13	CATS2D_01_AL	5	nCp	13	Mor30s
6	SpMAD_B(s)	14	Mor26v	6	R1v+	14	ChiA_X
7	Mor21s	15	Mor27u	7	H-052	15	MATS3v
8	AAC	16	MATS3e	8	Mor21s		
AS11HC - *a* values				**AS11HC - *b* values**			
1	nO	11	Eta_betaS_A	1	nO	11	Mor26s
2	MW	12	S3K	2	X3Av	12	Eig04_EA(dm)
3	Mi	13	Eig08_AEA(dm)	3	G2u	13	MATS3i
4	nF	14	Psi_i_0d	4	Mor26v	14	SM12_AEA(bo)
5	Eta_F_A	15	MATS2e	5	RDF035s	15	Mor28m
6	Eig09_AEA(bo)	16	E1u	6	E1s	16	Mor10m
7	SPH			7	SPH	17	RDF055m
8	nR=Cp			8	nF	18	MATS5i
9	J_Dz(i)			9	J_Dz(i)	19	HOMA
10	E3u			10	Mor09m	20	ASP

Table 4.6 Coefficients and their standard errors of QSRR models on the AS20 column.

***a* values**			***b* values**		
Descriptor	**Coefficient**	**Standard error**	**Descriptor**	**Coefficient**	**Standard error**
Mor15m	0.077	0.050	Mor30m	0.018	0.019
MATS2e	-0.086	0.044	Mor21s	0.119	0.018
MATS4m	-0.031	0.031	nCconj	-0.042	0.013
nO	-0.672	0.039	Mor17s	-0.091	0.016
CATS2D_01_AL	0.467	0.042	E1u	-0.029	0.019
GATS6p	0.078	0.034	nN	-0.051	0.017
SpMax_AEA(dm)	0.243	0.039	nF	0.183	0.016
nN	0.287	0.043	Mor32i	0.125	0.019
Eig09_AEA(bo)	-0.143	0.034	L3m	0.045	0.020
MW	0.362	0.051	RDF035s	-0.160	0.021
E2s	0.019	0.038	nO	0.655	0.024
Mor17m	0.123	0.047	CATS2D_01_AL	-0.320	0.014
P_VSA_s_3	-0.202	0.075	SpMax_AEA(dm)	-0.198	0.024
H-047	0.100	0.039	RDF020e	-0.160	0.018
E2p	-0.068	0.045	SpMAD_B(s)	-0.026	0.019

MEcc	-0.066	0.040	RARS	0.051	0.023
SpMin7_Bh(i)	0.109	0.028	MATS2e	0.034	0.020
MATS8p	-0.046	0.024	E3p	-0.059	0.018
Mor04m	0.052	0.039	E3u	-0.001	0.024
Mor10e	-0.151	0.057	ATS8m	-0.060	0.018
Eig07_EA(dm)	-0.125	0.044	Mor29u	-0.048	0.024
Mor25i	-0.030	0.036	H-051	-0.028	0.020
Intercept	-2.332	0.023	Intercept	1.461	0.010

Table 4.7 Coefficients and their standard errors of QSRR models on the AS19 column.

***a* values**			***b* values**		
Descriptor	**Coefficient**	**Standard error**	**Descriptor**	**Coefficient**	**Standard error**
H0i	-0.100	0.021	P_VSA_s_3	0.351	0.014
SIC0	0.055	0.031	LDI	0.250	0.015
MATS6p	-0.024	0.024	GATS6e	-0.035	0.012
E2s	0.099	0.019	Mor09e	-0.173	0.010
nO	-0.214	0.033	nCp	-0.079	0.011
SpMAD_B(s)	-0.182	0.023	R1v+	0.088	0.010

Mor21s	-0.357	0.021	H-052	0.084	0.010
AAC	0.193	0.026	Mor21s	0.131	0.017
MEcc	-0.415	0.027	Eig04_EA(dm)	0.111	0.012
nCconj	0.098	0.017	Mor27u	0.082	0.014
MATS6v	0.032	0.022	RDF035v	-0.043	0.013
SaasC	0.117	0.017	PDI	-0.063	0.012
CATS2D_01_AL	0.194	0.020	Mor30s	-0.025	0.016
Mor26v	-0.122	0.021	ChiA_X	-0.067	0.012
Mor27u	-0.021	0.021	MATS3v	0.014	0.008
MATS3e	0.082	0.019	Intercept	1.324	0.008
Intercept	-1.884	0.014			

Table 4.8 Coefficients and their standard errors of QSRR models on the AS11HC column.

***a* values**			***b* values**		
Descriptor	**Coefficient**	**Standarderror**	**Descriptor**	**Coefficient**	**Standard error**
nO	-1.259	0.040	nO	0.518	0.020
MW	0.865	0.047	X3Av	-0.047	0.014
Mi	0.509	0.049	G2u	0.167	0.013

nF	-0.553	0.039	Mor26v	0.147	0.021
Eta_F_A	0.365	0.052	RDF035s	-0.148	0.014
Eig09_AEA(bo)	-0.101	0.020	E1s	-0.142	0.018
SPH	0.210	0.040	SPH	-0.139	0.025
nR=Cp	-0.072	0.018	nF	0.104	0.014
J_Dz(i)	0.440	0.047	J_Dz(i)	-0.324	0.025
E3u	0.129	0.033	Mor09m	-0.083	0.012
Eta_betaS_A	-0.084	0.030	Mor26s	-0.079	0.018
S3K	0.068	0.033	Eig04_EA(dm)	0.120	0.019
Eig08_AEA(dm)	-0.110	0.027	MATS3i	-0.036	0.012
Psi_i_0d	-0.058	0.021	SM12_AEA(bo)	-0.101	0.028
MATS2e	-0.163	0.025	Mor28m	-0.090	0.021
E1u	-0.022	0.022	Mor10m	-0.016	0.012
Intercept	-2.075	0.018	RDF055m	0.064	0.022
			MATS5i	0.025	0.014
			HOMA	-0.037	0.015
			ASP	-0.063	0.026
			Intercept	1.353	0.010

where y_i are the measured *a* and *b* values, $\hat{y}_i$ are the predicted *a* and *b* values, and *n* is the total number of ions used in the training set.

To evaluate whether these constructed QSRR models could robustly predict the *a* and *b* values for new ions that are not involved in the model generation, external validation was performed on external test sets which consist of 5, 4 and 4 ions on the AS20, AS19 and AS11HC columns, respectively, *i.e.*, 10% of the full data sets **(Table 4.1)**. The external test ions were chosen with them havin as many similar ions (based on Tanimoto similarity index which will be described in detail in Chapter 5) in the training set as possible and having different charges (*i.e.*, monovalent and divalent ions). Compared to internal validation, such as cross-validation, external validation is a more reliable method to estimate the predictive power of the models since it is testing completely new compounds [33]. Three different validation functions for the predictive squared correlation coefficient Q^2 were assessed:

$$R^2_{pre} = 1 - \frac{\sum_i^m (y_{i,t} - \hat{y}_{i,t})^2}{\sum_i^m (y_{i,t} - \bar{y})^2} \quad \textbf{(4.6)}$$

$$Q^2_{ext(F2)} = 1 - \frac{\sum_i^m (y_{i,t} - \hat{y}_{i,t})^2}{\sum_i^m (y_{i,t} - \overline{y_t})^2} \quad \textbf{(4.7)}$$

$$Q^2_{ext(F3)} = 1 - \frac{\left[\sum_i^m (y_{i,t} - \hat{y}_{i,t})^2\right]/m}{\left[\sum_i^n (y_i - \bar{y})^2\right]/n} \quad \textbf{(4.8)}$$

where R_{pre}^2 (or $Q_{ext(F1)}^2$) is a commonly used function for external validation, $y_{i,t}$ are the measured values for test set ions, $\hat{y}_{i,t}$ are the predicted values for test set ions, $\bar{y}$ is the mean measured value for training set ions, $\overline{y_t}$ is the mean measured value for test set ions, y_i are the measured values for training set ions, m is the total number of ions in the external test set and n is the total number of ions in the training set [6, 34, 35].

Additionally, the root mean square error of prediction (*RMSEP*), *i.e.*, the average prediction error for the external test set, was calculated by:

$$RMSEP = \left[\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2\right]^{0.5} \quad \textbf{(4.9)}$$

where y_i are the measured *a* and *b* values for external test set ions, $\hat{y}_i$ are the predicted *a* and *b* values for external test set ions, and n_{ext} is the total number of ions used in the external test set.

Table 4.9 summarises the statistical and validation parameters indicating the predictive performance for the six QSRR models built in this study. Appropriate models with good fit and high accuracy were obtained, with R^2 >0.98 and *RMSE* <0.11 for all six models. In addition, the resultant *F* and *p* values confirm that these models are statistically significant [5, 10, 36]. $Q_{ext(Fn)}^2$ values were greater than 0.8 except for the model for *a* values on the AS11HC column. According to some recent reports [34, 37] regarding the evaluation of external validation in quantitative structure-activity relationship (QSAR) models, $Q_{ext(F3)}^2$ can be used as a robustness criterion. $Q_{ext(F3)}^2$ has the advantage of being independent of the size and distribution of external test sets, unlike $Q_{ext(F1)}^2$ and $Q_{ext(F2)}^2$. The $Q_{ext(F3)}^2$ values for the developed QSRR models were greater than 0.93 (except for the model for *a* values on AS11HC column, $Q_{ext(F3)}^2$ = 0.78). Considering $Q_{ext(F1)}^2$ and $Q_{ext(F2)}^2$ values can suffer from possible biases due to small-sized external test sets, it is more desirable to use $Q_{ext(F3)}^2$ as a measure of evaluation for external validation in this study, which employs external test sets. As a result, all six models showed an acceptable level of $Q_{ext(F3)}^2$ values greater than 0.7 [37], along with low *RMSEP* values < 0.4.

Further, the developed models were evaluated by using other statistical criteria designed to assess predictivity of models [20, 33], namely: (i) R^2 >0.6; (ii) $(R^2 - R_0^2)/R^2$ <0.1 and 0.85 < s_0 < 1.15; (iii) $\left|R_0^2 - R'^2_0\right| < 0.3$, where R^2 is the coefficient of determination between the predicted and measured values for the external test set, R_0^2 and R'^2_0 are, respectively, the coefficients of determination for predicted *vsersus* measured values and measured *versus* predicted values for regression through the origin for the external test set, and s_0 is slope of the regression line through the origin

Table 4.9 Predictive performance of QSRR models on various columns.

Models	Training				External test			
	R^2	*RMSE*	*F*	*p*-value	R_{pre}^2	$Q_{ext(F2)}^2$	$Q_{ext(F3)}^2$	*RMSEP*
AS20-*a*	0.984	0.106	59	2.92E-14	0.928	0.866	0.954	0.181
AS20-*b*	0.994	0.044	147	2.42E-18	0.930	0.919	0.939	0.136
AS19-*a*	0.990	0.059	126	1.49E-16	0.869	0.804	0.929	0.159
AS19-*b*	0.995	0.033	270	1.48E-20	0.925	0.924	0.931	0.119
AS11HC-*a*	0.993	0.071	170	4.45E-17	0.687	0.618	0.784	0.397
AS11HC-*b*	0.996	0.031	200	1.00E-14	0.915	0.914	0.930	0.134

for the external test set. All models except for the QSRR model for *a* values on the AS11HC column satisfied all these criteria (R^2 >0.9, 0.9 < s_0 < 1.1, $(R^2 - R_0^2)/R^2$ <0.06, $\left|R_0^2 - R'^2_0\right| < 0.05$). However, the QSRR model for the *a* values on the AS11HC column was not successful for the criteria $(R^2 - R_0^2)/R^2$ and $\left|R_0^2 - R'^2_0\right|$, due to low values of R'^2_0. These results correspond to the low $Q_{ext(Fn)}^2$-values obtained on this column. It should be mentioned that this model did satisfy the other evaluation criteria (R^2 value and k range).

Figure 4.2 illustrates correlations between the measured and the predicted values (*a* and *b*) for the training and external test sets. In addition to the statistical and validation parameters in **Table 4.9**, all six QSRR models generally showed good fit to the 45 degree line (dotted line in **Fig. 4.2A-F**), further indicating the good predictive ability of the constructed models [37]. However, one data point among external test ions in the **Fig. 2E** (*a*-values on AS11HC column) showed a relatively higher deviation from the 45 degree line compared to the other models, corresponding to the lower $Q_{ext(F3)}^2$ and higher *RMSEP* values (0.78 and 0.40 respectively), as well as high values of $(R^2 - R_0^2)/R^2$ and $\left|R_0^2 - R'^2_0\right|$. There are clusters of data points in **Figs 4.2B**, **1D**, and **1F** which correspond to monovalent, divalent, and trivalent ions, since it is the charge on the analyte ion which determines the *b* values in the LSS model equation.

4.3.3. Application of QSRR models for *a* and *b* values to the prediction of retention times

To ascertain whether the generated QSRR models for *a* and *b* values can be used for accurate prediction of retention times for anions, the developed QSRR models for *a* and *b* values were applied to the linear solvent strength (LSS) model **(Eq. 4.1)**. Specifically, the predicted *a* and *b* values for each anion, obtained using the constructed QSRR models, and different eluent concentrations (35 mM on AS20; 25 mM on AS19;

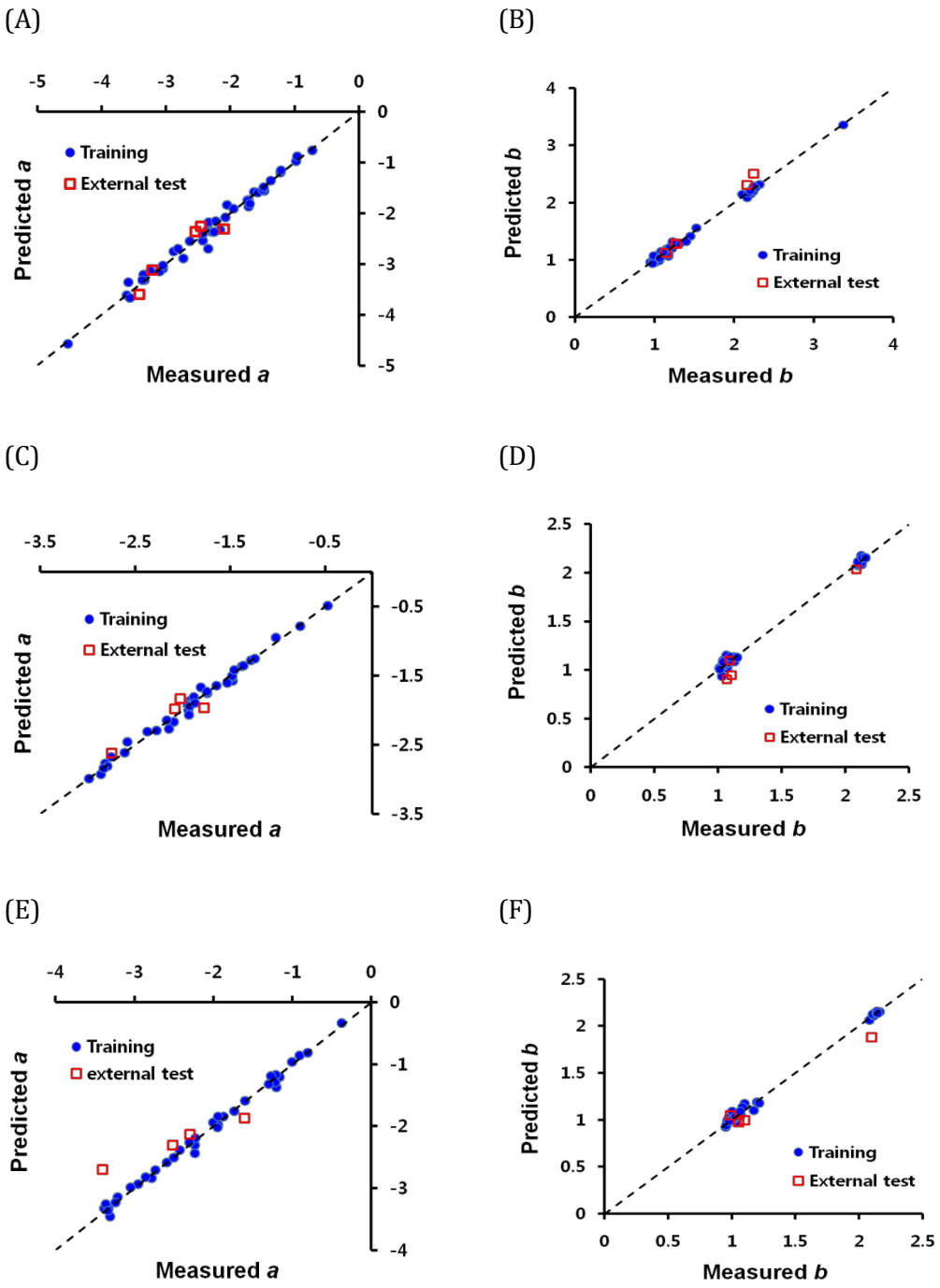


Figure 4.2 Correlation between predicted and measured values. The QSRR models for *a* values on the AS20 column (A) and *b* values (B), *a* values on the AS19 column (C) and *b* values (D), and *a* values on the AS11HC column (E) and *b* values (F) were used to calculated the predicted values.

30 mM on AS11HC) were substituted into **Eq. 4.1** to calculate retention times at the corresponding eluent concentration on each column. The predicted retention times of the anions used in this study (on three columns at three eluent concentrations) were compared with their corresponding measured retention times. **Figure 4.3** shows a plot of the predicted *versus* measured retention times for the training and external test sets, and includes a total of 130 data points (117 in the training set sand 13 in the external test sets). A strong correlation between the predicted and measured retention times was found with an R^2-value of 0.98 and an *RMSE* of 0.89 min. Additionally, the $Q_{ext(F3)}^2$ and *RMSEP* values for the external test ions were 0.96 and 1.18 min, respectively. It can therefore be concluded that the QSRR models for *a* and *b* values in the LSS model **(Eq. 4.1)** can be used successfully to predict the retention times of inorganic and small organic anions in IC based only on their structures, with good accuracy and predictive ability.

4.4. Conclusions

In this study QSRR models for both *a* and *b* values in the LSS model in IC were successfully generated for around 45 inorganic and small organic anions on three different columns (AS20, AS19 and AS11HC). The optimal subset of molecular descriptors used for building each QSRR model was determined in two steps: first, selection of the optimal number of descriptors corresponding to the highest correlation (R^2 >0.99) and the smallest RMSE, followed by elimination of multi-collinear descriptors. QSRR models were constructed using MLR by correlating retention parameters (*a* and *b* values) to the selected descriptors. High accuracy models were developed for both the *a* and *b* values on the three columns (AS20, AS19 and AS11HC). External validation was performed using 10% of the total data set, resulting in acceptable $Q_{ext(F3)}^2$ values above 0.7 for all six models, supporting the potential of QSRR

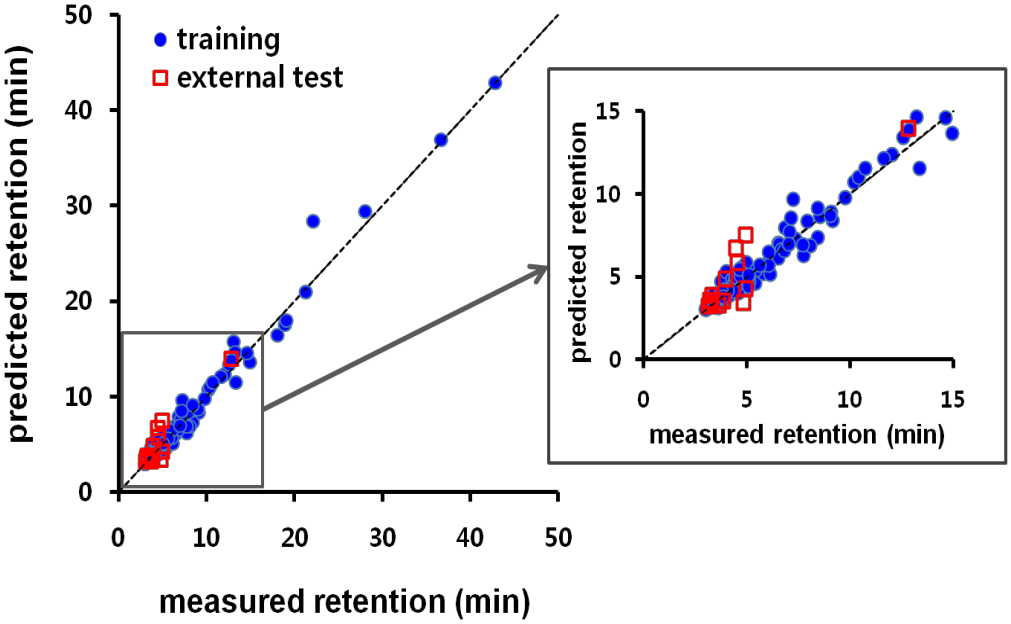


Figure 4.3 Predicted retention times versus measured retention times. A total of 130 data points (117 in the training sets and 13 in the external test sets) on the three columns (AS20, AS19 and AS11HC) are included on the graph. The retention times obtained at 35 mM on AS 20 column, 25 mM on AS19 and 30 mM on AS11HC were used.

models as predictive tools in IC. Furthermore, the predicted *a* and *b* values can be used to predict the retention times of analytes under isocratic conditions, with an acceptable level of accuracy and predictive ability (average R^2 of 0.98; *RMSE* of 0.89 min; $Q_{ext(F3)}^2$ of 0.96; *RMSEP* of 1.18 min). This approach can easily be extended to the prediction of retention times under gradient conditions and with multi-step eluent profiles. Additionally, the approach can be applicable to any type of IC column and eluent for which appropriate isocratic retention data can be acquired.

It is concluded that the employed QSRR approach can enable the prediction of retention times of unknown inorganic and small anions over a broad range of eluent concentrations on a wide range of columns. Therefore, retention prediction via molecular descriptors and the LSS model can lead to faster and more robust method development in IC.

4.5. References

[1] R. Kaliszan, QSRR: Quantitative Structure-(Chromatographic) Retention Relationships, Chem. Rev. 107 (2007) 3212-3246.

[2] K. Heberger, Quantitative structure-(chromatographic) retention relationships, J. Chromatogr. A 1158 (2007) 273-305.

[3] G. Carlucci, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure-retention relationships, Anal. Chim. Acta 601 (2007) 68-76.

[4] J. Ghasemi, S. Saaidpour, QSRR prediction of the chromatographic retention behavior of painkiller drugs, J. Chromatogr. Sci. 47 (2009) 156-163.

[5] K. Gorynski, B. Bojko, A. Nowaczyk, A. Bucinski, J. Pawliszyn, R. Kaliszan, Quantitative structure-retention relationships models for prediction of high

performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds, Anal. Chim. Acta 797 (2013) 13-19.

[6] N. Kritikos, A. Tsantili-Kakoulidou, Y.L. Loukas, Y. Dotsikas, Liquid chromatography coupled to quadrupole-time of flight tandem mass spectrometry based quantitative structure-retention relationships of amino acid analogues derivatized via n-propyl chloroformate mediated reaction, J. Chromatogr. A 1403 (2015) 70-80.

[7] C.B. Mazza, N. Sukumar, C.M. Breneman, S.M. Cramer, Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure, Anal. Chem. 73 (2001) 5457-5461.

[8] M. Song, C.M. Breneman, J. Bi, N. Sukumar, K.P. Bennett, S. Cramer, N. Tugcu, Prediction of protein retention times in anion-exchange chromatography systems using support vector regression, J. Chem. Inf. Comput. Sci. 42 (2002) 1347-1357.

[9] G. Malmquist, U.H. Nilsson, M. Norrman, U. Skarp, M. Stromgren, E. Carredano, Electrostatic calculations and quantitative protein retention models for ion exchange chromatography, J. Chromatogr. A 1115 (2006) 164-186.

[10] S. Studzinska, M. Molikova, P. Kosobucki, P. Jandera, B. Buszewski, Study of the Interactions of Ionic Liquids in IC by QSRR, Chromatographia 73 (2011) 35-44.

[11] Š. Ukić, M. Novak, P. Žuvela, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part I: Conventional QSRR approach, Chromatographia 77 (2014) 985-996.

[12] Š. Ukić, M. Novak, A. Vlahović, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part II: Artificial intelligence QSRR approach, Chromatographia 77 (2014) 997-1007.

[13] V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, Prediction of capillary gas chromatographic retention times of fatty acid methyl

esters in human blood using MLR, PLS and back-propagation artificial neural networks, Talanta 83 (2011) 1014-1022.

[14] J. Yan, D.S. Cao, F.Q. Guo, L.X. Zhang, M. He, J.H. Huang, Q.S. Xu, Y.Z. Liang, Comparison of quantitative structure-retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds, J. Chromatogr. A 1223 (2012) 118-125.

[15] A.G. Fragkaki, A. Tsantili-Kakoulidou, Y.S. Angelis, M. Koupparis, C. Georgakopoulos, Gas chromatographic quantitative structure-retention relationships of trimethylsilylated anabolic androgenic steroids by multiple linear regression and partial least squares, J. Chromatogr. A 1216 (2009) 8404-8420.

[16] A.G. Fragkaki, E. Farmaki, N. Thomaidis, A. Tsantili-Kakoulidou, Y.S. Angelis, M. Koupparis, C. Georgakopoulos, Comparison of multiple linear regression, partial least squares and artificial neural networks for prediction of gas chromatographic relative retention times of trimethylsilylated anabolic androgenic steroids, J. Chromatogr. A 1256 (2012) 232-239.

[17] S. Schefzick, C. Kibbey, M.P. Bradley, Prediction of HPLC conditions using QSPR techniques: an effective tool to improve combinatorial library design, J. Comb. Chem. 6 (2004) 916-927.

[18] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemom. Intell. Lab. Syst. 76 (2005) 185-196.

[19] M. Goodarzi, R. Jensen, Y. Vander Heyden, QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions, J. Chromatogr. B 910 (2012) 84-94.

[20] M. Talebi, G. Schuster, R.A. Shellie, R. Szucs, P.R. Haddad, Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modeling of retention times in reversed-phase liquid chromatography, J. Chromatogr. A 1424 (2015) 69-76.

[21] P. Zuvela, J.J. Liu, K. Macur, T. Baczek, Molecular descriptor subset selection in theoretical peptide quantitative structure-retention relationship model development using nature-inspired optimisation algorithms, Anal. Chem. 87 (2015) 9876-9883.

[22] B. Law, S. Weir, Quantitative structure-retention relationships for secondary interactions in cation-exchange liquid chromatography, J. Chromatogr. A 657 (1993) 17-24.

[23] C.B. Mazza, C.E. Whitehead, C.M. Brenernan, S.M. Crarner, Predictive Quantitative Structure Retention Relationship Models for Ion-Exchange Chromatography, Chromatographia 56 (2002) 147-152.

[24] P.E. Morgan, D.J. Barlow, M. Hanna-Brown, R.J. Flanagan, Artificial neural network modeling of the retention of acidic analytes in strong anion-exchange HPLC: Elucidation of structure-retention relationships, Chromatographia 75 (2012) 693-700.

[25] B.K. Ng, R.A. Shellie, G.W. Dicinoski, C. Bloomfield, Y. Liu, C.A. Pohl, P.R. Haddad, Methodology for porting retention prediction data from old to new columns and from conventional-scale to miniaturised ion chromatography systems, J. Chromatogr. A 1218 (2011) 5512-5519.

[26] R.A. Shellie, B.K. Ng, G.W. Dicinoski, S.D.H. Poynter, J.W. O'Reilly, C.A. Pohl, P.R. Haddad, Prediction of analyte retention for ion chromatography separations performed using elution profiles comprising multiple isocratic and gradient steps, Anal. Chem. 80 (2008) 2474-2482.

[27] S.H. Park, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl, Enhanced methodology for porting ion chromatography retention data, J. Chromatogr. A 1436 (2016) 59-63.

[28] R. Leardi, Genetic algorithms in chemistry, J. Chromatogr. A 1158 (2007) 226-233.

[29] Y.K. Tu, M. Kellett, V. Clerehugh, M.S. Gilthorpe, Problems of correlations between explanatory variables in multiple regression analyses in the dental literature, British Dental J. 199 (2005) 457-461.

[30] J.F.J. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, Multivariate Data Analysis, 3rd ed., Macmillan, New York, 1995.

[31] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions, J. Mol. Graph. Model. 23 (2005) 503-523.

[32] J. Ghasemi, S. Saaidpour, QSRR Prediction of the chromatographic retention behavior of painkiller drugs, J. Chromatogr. Sci. 47 (2009) 156-163.

[33] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: Validation in the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69-77.

[34] V. Consonni, D. Ballabio, R. Todeschini, Comments on the definition of the Q2 parameter for QSAR validation, J. Chem. Inf. Model. 49 (2009) 1669-1678.

[35] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, J. Chemometr. 24 (2010) 194-201.

[36] T. Baczek, R. Kaliszan, Combination of linear solvent strength model and quantitative structure-retention relationships as a comprehensive procedure of approximate prediction of retention in gradient liquid chromatography, J. Chromatogr. A 962 (2002) 41-55.

[37] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for

scatter plot inspection, J. Chem. Inf. Model. 52 (2012) 2044-2058.

Chapter 5

Structural similarity-based QSRR modelling

5.1. Introduction

Method development in liquid chromatography (LC) largely consists of two stages - "scoping" and "optimisation". The "scoping" stage involves the estimation of general chromatographic parameters, which include the chromatographic mode to be used (*e.g.*, reversed-phase liquid chromatography [RPLC], hydrophilic interaction liquid chromatography [HILIC] or ion chromatography [IC]), stationary phase, broad mobile phase composition, detection mode, *etc.* The "optimisation" stage is comprised of establishing the detailed chromatographic conditions, such as the exact composition of the mobile phase, pH, temperature, and the flow-rate. Computer-assisted method development is recognised as a powerful alternative to the trial-and-error approach in LC method development, which is often laborious, time-consuming, and expensive. By employing retention modelling and chromatographic simulation, the selection of chromatographic parameters and conditions is facilitated, resulting in rapid, robust and rugged method development.

In IC, computer-aided approaches have been mainly utilised for streamlining the optimisation phase, rather than the scoping phase [1-6]. For this purpose, both isocratic retention models (such as the Linear Solvent Strength [LSS] model and LSS model-empirical approach [LSSM-EA] for eluents including two competing ions) as well as gradient retention models (such as the Rocklin model [7], the Jandera model [8], based on the isocratic LSS model, and the quadratic linear solvent strength gradient model) can be used [1]. A commercially available example is the Virtual Column® software, marketed by Thermo Fisher Scientific, which is now used routinely for IC optimisation method development. The isocratic LSS model is a representative model

embedded in this software and, for a given temperature, is expressed as:

$$\log k = a - b\log[\mathrm{OH}^-] \quad \textbf{(5.1)}$$

where k is the retention factor, $[\mathrm{OH}^-]$ is hydroxide concentration in mol/L, and a- and b-values are the intercept and slope, respectively [9-11]. Virtual Column® includes the a- and b-values for each analyte in its database. Retention times of the analytes can then be predicted for all eluent compositions under isocratic, as well as gradient and multi-step elution conditions [10].

For streamlining the scoping phase of method development, Quantitative Structure-Retention Relationships (QSRR) modelling can be a powerful approach [12] since QSRR enables the prediction of retention based only on the chemical structures of analytes [13], leading to reductions in the time and cost of the method development process by minimising the number of experiments. In addition to the retention prediction of unknown compounds [14-17], QSRR modelling has been also used for the classification and characterisation of stationary phases [18-20], the elucidation of retention mechanisms [19, 21, 22], and the identification of challenging unknown metabolites, when combined with LC-high resolution mass spectrometry [23].

In QSRR studies, molecular modelling is often used to calculate a large number (typically >3000) of molecular descriptors for each analyte under study, followed by the implementation of a suitable variable selection technique to identify those molecular descriptors which best describe the retention of the analytes [24-27]. Among many variable selection techniques being successfully employed in QSRR, the genetic algorithm (GA) which was used in this study is well accepted as one of the most powerful methods [16, 28]. Finally, a variety of modelling techniques, including multiple linear regression (MLR) [14, 15, 29, 30], partial least square (PLS) regression [12, 15, 26, 29, 31], support vector machine (SVM) [32], and artificial neural network (ANN) [16, 22, 29], have been extensively employed for generating the QSRR models

from the selected descriptors. While MLR is the most commonly used modelling technique in QSRR because MLR is a simple regression method that is easy to interpret [29], MLR is also known to be vulnerable to chance correlation, over-fitting, and multi-co-linearity of descriptors [26]. These problems can be minimized using PLS regression, where the dependent variable (retention parameter) is correlated indirectly to independent variables (descriptors matrix) *via* a small number of linear and orthogonal latent variables (LVs). The optimum number of LVs is obtained *via* cross-validation to maximize the covariance between the descriptor matrix and the retention parameter. Unlike MLR, PLS models can conveniently handle the most common scenario in QSRR studies where the number of descriptors is much greater than the number of analytes in the training set [26, 29, 31].

It seems reasonable to expect more prediction accuracy from a QSRR model if the model is trained with a dataset that contains only compounds which are similar to the target compound under study. This strategy, which was initially introduced in Quantitative Structure Activity Relationship (QSAR) modelling [33, 34], underlines the "Similar Property Principle", stating that structurally similar molecules tend to have similar properties [35]. For example, according to a QSAR study using the fathead minnow dataset containing 322 organic compounds, compounds more similar to a query compound resulted in more reliable prediction of a query compound's acute toxicity, where a hierarchical clustering analysis was used to incude similar compounds in the training set [33]. In the same way, a proof-of-concept study for the applicability of this strategy in QSRR modelling has been reported [12, 36]. This approach of developing individual local models for each compound under study and then combining the predictions from these models to predict an overall chromatogram is known as the "federation of local models". Tyteca *et al.* [12] and Talebi *et al.* [36] have reported that the prediction accuracy of local PLS models was improved significantly

by employing as a training subset the 20 most similar compounds to each query compound, where similar compounds were identified by ranking them based on their pair-wise Tanimoto Similarity (TS) score to the query compound [36].

In this chapter, the role of structural similarity of compounds in QSRR modelling was investigated in more detail in ion-exchange chromatography using a much broader range of anionic and cationic analytes, with a focus on two significant influencing factors: the extent of the similarity of compounds in the training set, expressed as the average TS score, and the neighbour count (NC) (the number of similar compounds in the training set). Accordingly, QSRR models were generated using training sets that had only similar neighbours to the target ion, based on different TS scores as the similarity threshold. Two retention parameters (*a*- and *b*-values in **Eq. 5.1**) in the LSS model were predicted using the QSRR models. The QSRR models were developed by a GA-PLS method and assessed by both internal cross-validation and external validation, mainly employing measures recommended by Tropsha *et al.* [37] for reliable model predictability. The applicability of QSRR models for the *a*- and *b*-values in retention time prediction was evaluated by plotting the experimental retention times against the predicted retention times obtained from fitting the predicted *a*- and *b*-values into the LSS model equation **(Eq. 5.1)**. This work was performed on databases containing both small anionic species (consisting primarily of inorganic anions) and larger organic cationic species of pharmaceutical interest.

5.2. Materials and Methods

5.2.1. Datasets

The anion datasets consisted of the *a*- and *b*-values for 36, 31 and 28 monovalent ions on Thermo Fisher Scientific IonPac AS20, AS19 and AS11HC columns (4 mm I.D. × 250 mm), respectively **(Table 5.1)**. The ranges of molecular mass of these ions on AS20,

Table 5.1 Anion datasets consisting of *a* and *b* values of monovalent ions on AS20, AS19 and AS11HC columns.

	AS20		AS19		AS11HC	
Compound	***a***	***b***	***a***	***b***	***a***	***b***
acetate	-2.63	1.32	-2.09	1.11	-2.73	1.18
acrylate	-2.33	1.22	-1.92	1.09	-2.23	1.05
benzenesulfonate	-1.37	1.01	-	-	-	-
benzoate	-1.47	1.02	-1.28	1.02	-1.00	0.96
bromoacetate	-2.05	1.13	-1.74	1.07	-1.73	1.00
butanesulfonate	-2.10	1.14	-1.78	1.07	-1.61	0.99
chlorate	-1.63	1.07	-1.38	1.05	-1.16	1.00
chloroacetate	-2.16	1.17	-1.81	1.07	-1.94	1.04
dibromoacetate	-1.57	1.04	-1.36	1.03	-0.91	0.97
dichloroacetate	-1.71	1.06	-1.48	1.04	-1.20	0.98
difluoroacetate	-2.07	1.15	-1.75	1.07	-1.87	1.02
ethanesulfonate	-2.37	1.23	-1.94	1.09	-2.22	1.07
fluoroacetate	-2.46	1.27	-	-	-	-
formate	-2.43	1.27	-1.94	1.09	-2.43	1.10
glycolate	-2.88	1.45	-2.15	1.13	-2.78	1.20
heptanesulfonate	-1.49	1.00	-	-	-	-
hexanesulfonate	-1.71	1.04	-1.48	1.03	-0.81	0.95
lactate	-2.81	1.39	-2.17	1.13	-2.85	1.22

methacrylate	-2.23	1.17	-1.88	1.08	-2.00	1.02
methanesulfonate	-2.30	1.19	-1.94	1.10	-2.30	1.09
n-butyrate	-2.43	1.22	-2.03	1.09	-2.29	1.05
nitrate	-1.48	1.06	-1.24	1.05	-1.21	1.00
nitrite	-1.73	1.08	-1.48	1.05	-1.59	1.01
n-valerate	-2.27	1.17	-1.93	1.08	-1.94	1.00
octanesulfonate	-1.22	0.95	-	-	-	-
p-chlorobenzenesulfonate	-0.98	0.98	-	-	-	-
pentanesulfonate	-1.95	1.11	-1.65	1.05	-1.21	0.97
perchlorate	-0.72	1.02	-0.47	1.04	-	-
propanesulfonate	-2.25	1.19	-1.87	1.08	-1.95	1.03
propionate	-2.54	1.27	-2.08	1.11	-2.51	1.10
pyruvate	-2.35	1.23	-1.93	1.09	-2.23	1.07
quinate	-3.10	1.53	-2.27	1.15	-	-
sorbate	-1.73	1.04	-1.53	1.03	-1.27	0.97
tribromoacetate	-0.96	0.98	-0.76	1.01	-	-
trichloroacetate	-1.21	0.99	-1.02	1.02	-0.37	0.96
trifluoroacetate	-1.70	1.06	-1.46	1.04	-1.30	0.98

AS19 and AS11HC were from 45 (formate) to 296 (tribromoacetate), 45 (formate) to 296 (tribromoacetate), and 45 (formate) to 217 (dibromoacetate), respectively. Ions in the training set were clustered based on their charges – a reasonable method, especially for the modelling of the *b*-values which represent the ratio of charge of analyte ions to the eluent ion. As the numbers of similar divalent and trivalent ions in the original databases were not enough to make clusters with a reasonable size (*i.e.*, at least 7), the present study was limited to modelling only monovalent ions. The *a*- and *b*-values were obtained from Virtual Column software and recalibrated by the previously reported "porting" methodology (Chapter 3) [9]. Conditions for separation included an eluent flow-rate of 1.0 mL/min, a column temperature of 30°C, and suppressed conductivity detection operated at 35°C, for details please see Chpater 3 [9].

The cation dataset **(Table 5.2)** included 87 larger molecular weight organic monovalent ions (molecular mass from 32 [methylamine] up to 506 [dipyridamole]). Retention data for these compounds were obtained experimentally on a Dionex (Sunnyvale, CA, USA) ICS-3000 Ion Chromatography system, comprising a dual gradient pump unit (DP), a dual eluent generator unit (EG), dual suppressed conductivity detector compartment (DC), variable wavelength UV detector (VWD) and autosampler (AS). A CS17 analytical column (2 mm i.d. × 250 mm) with a CS17 guard column (2 mm i.d. × 50 mm) was used at 30 °C. The injection volume was 10 μL and the eluent flow-rate was 0.25 mL/min. Methanesulfonic acid (MSA) of various concentrations (see below) was pumped at 0.16 mL/min and mixed with acetonitrile (ACN) at 0.09 mL/min through a T-piece connector followed by a gradient mixer (Dionex GM-4 2mm). An additional pump (Jasco PU-2089i Plus, Tokyo, Japan) was used to supply water at 0.5 mL/min to the EG, continuously regenerated trap column (CR-TC) and degasser. The retention times were collected for 5 eluent compositions under

Table 5.2 Cation dataset consisting of *a* and *b* values of monovalent ions on the CS17 column.

ID	Compound	*a*	*b*	R^2	Detection
1	1-3-Methoxyphenyl -2-methylaminoethanol	-1.78	0.98	1.000	CD
2	1-Butylamine	-2.08	1.04	1.000	CD
3	1-Methyl-3-phenyl propylamine	-1.55	0.95	1.000	CD
4	2-2-Aminoethoxyethanol	-2.42	1.12	0.999	CD
5	2-Amino-1-phenylethanol	-1.93	1.00	1.000	CD
6	2-Diethylaminoethanol	-2.09	1.04	1.000	CD
7	2-Dimethylaminoethanol	-2.21	1.07	1.000	CD
8	2-Ethylaminoethanol	-2.29	1.09	1.000	CD
9	2-Methylaminoethanol	-2.36	1.11	1.000	CD
10	3-Amino-1-propanol	-2.46	1.14	1.000	CD
11	3-Methoxytyramine	-2.04	0.97	0.999	UV
12	3-Methylphenethylamine	-1.65	0.92	0.998	UV
13	3-Phenylpropylamine	-1.61	0.96	1.000	CD
14	4-Epitetracycline	-1.57	0.90	0.998	UV
15	4-Phenylbutylamine	-1.48	0.94	1.000	CD
16	5-Amino-1-pentanol	-2.32	1.09	1.000	CD
17	Acebutolol	-1.63	0.91	0.998	UV
18	Alprenolol	-1.26	0.88	0.998	UV
19	Amino-2-propanol	-2.45	1.14	0.999	CD
20	Amoxicillin	-2.09	0.92	0.996	UV
21	Atenolol	-1.88	0.94	0.999	UV
22	Betaxolol	-1.25	0.89	0.998	UV
23	Bethanechol	-1.97	1.02	1.000	CD
24	Bisoprolol	-1.44	0.90	0.998	UV
25	Carbachol	-2.02	1.03	1.000	CD
26	Carvedilol	-0.73	0.89	0.999	UV
27	Celiprolol	-1.46	0.90	0.998	UV

28	Chlortetracycline	-1.28	0.88	0.998	UV
29	Choline	-2.07	1.04	1.000	CD
30	Cimetidine	-1.90	0.95	0.999	UV
31	Clenbuterol	-1.44	0.90	0.998	UV
32	Clomipramine	-0.54	0.90	0.999	UV
33	Clonidine	-1.47	0.91	0.998	UV
34	Clorprenaline	-1.58	0.91	0.998	UV
35	Cyclohexylamine	-1.89	1.00	1.000	CD
36	Diethanolamine	-2.49	1.15	0.997	CD
37	Dimethylamine	-2.30	1.10	1.000	CD
38	Diphenhydramine	-1.05	0.89	0.999	UV
39	Dipyridamole	-1.08	0.89	0.998	UV
40	Dopamine	-2.14	0.96	0.995	UV
41	Doxepin	-0.92	0.89	0.999	UV
42	Doxycycline	-1.26	0.87	0.997	UV
43	Esmolol	-1.53	0.90	0.998	UV
44	Ethanolamine	-2.52	1.16	0.999	CD
45	Etilefrine	-2.05	0.97	0.999	UV
46	Fenoterol	-2.02	0.95	0.999	UV
47	Hordenine	-1.88	0.94	0.999	UV
48	Hydroxyzine	-1.13	1.06	1.000	UV
49	Imipramine	-0.78	0.90	0.998	UV
50	Isoprenaline	-2.13	0.97	0.998	UV
51	Labetalol	-1.32	0.89	0.998	UV
52	Metanephrine	-2.10	0.97	0.999	UV
53	Metaproterenol	-2.13	0.95	0.994	UV
54	Methylamine	-2.46	1.15	0.999	CD
55	Metoprolol	-1.59	0.91	0.998	UV
56	Mexiletine	-1.50	0.91	0.998	UV
57	Morpholine	-2.18	1.07	1.000	CD
58	Nadolol	-1.83	0.93	0.998	UV
59	Nebivolol	-0.82	0.90	0.999	UV

60	Neostigmine	-1.60	0.91	0.998	UV
61	N-Methyldiethanolamine	-2.32	1.10	0.999	CD
62	N-Methylphenethylamine	-1.70	0.93	0.999	UV
63	N-Methylpyrrolidine	-1.96	1.02	1.000	CD
64	Norepinephrine	-2.19	0.94	0.992	UV
65	Norfenefrine	-2.13	0.96	0.996	UV
66	Normetanephrine	-2.15	0.97	0.998	UV
67	Octopamine	-2.14	0.96	0.995	UV
68	Oxprenolol	-1.43	0.90	0.998	UV
69	Oxytetracycline	-1.61	0.89	0.998	UV
70	Penbutolol	-0.78	0.84	0.998	UV
71	Phenethylamine	-1.79	0.94	0.999	UV
72	Phenylalanine	-1.84	0.83	0.999	UV
73	Phenylephrine	-2.11	0.98	0.999	UV
74	Pindolol	-1.54	0.89	0.998	UV
75	Promethazine	-0.81	0.90	0.999	UV
76	Propranolol	-1.14	0.88	0.997	UV
77	Propylamine	-2.23	1.07	1.000	CD
78	Pyrrobutamine	-0.37	0.86	0.997	UV
79	Salbutamol	-2.07	0.96	0.999	UV
80	Serotonin	-1.96	0.95	0.999	UV
81	Sulpiride	-1.79	0.93	0.999	UV
82	Synephrine	-2.12	0.97	0.999	UV
83	tert-Butylamine	-2.25	1.08	1.000	CD
84	Triethylamine	-1.94	1.01	1.000	CD
85	Trimethylamine	-2.14	1.05	1.000	CD
86	Tryptophan	-1.74	0.85	1.000	UV
87	Tyramine	-2.06	0.97	0.999	UV

isocratic conditions (5, 10, 20, 30 and 40 mM MSA with 36% ACN). Excellent correlation (R^2 >0.992) between log*k* *versus* log[MSA] was found for larger organic cations being studied, confirming the validity of the LSS model for the studied ions. This is due to the addition to the eluent of a relatively high content of organic solvent [38] and hence the minimisation of hydrophobic interactions [38]. Non-suppressed UV detection at 220 nm was utilised for the detection of 56 cations and 270 nm for synephrine and tyramin. The remaining 29 cations were detected by suppressed conductivity detection at 35°C, employing CSRS® 300 (2 mm) electrolytic suppressor. Instrument control and data acquisition were acquired using Chromeleon software (version 6.80).

5.2.2. Molecular descriptors

Molecular descriptors were calculated using the method in reference [12]. Briefly, 2D structures of compounds were drawn using MarvinSketch version 6.2.1 (ChemAxon, Budapest, Hungary) [39]. The 50 lowest energy 3D-conformers were obtained using the Merck Molecular Force Field (MMFF94) [40-43] (in Balloon [44, 45]). The lowest energy conformers were then geometrically optimised in water by the semi-empirical Parametric Method 7 (PM7) [46], performed in Molecular Orbital PACkage (MOPAC) [47]. The 3D structures for the resulting optimised geometries were uploaded into the Dragon 6.0 software [48] (Talete, Milano, Italy) for the calculation of molecular descriptors. From 4885 descriptors calculated initially, only 490 and 570 descriptors remained for the anion and cation datasets, respectively, following the removal of descriptors with constant values, with standard deviation ≤0.0001, or with at least one missing value, as well as descriptors with an absolute pair-wise correlation ≥0.9.

5.2.3. Training and test sets

Training sets were obtained by including ions similar to the target ion, based on

their pair-wise similarity values, *i.e.*, TS scores, to the target ion. JChem for Excel (ChemAxon, Budapest, Hungary) was used to calculate pair-wise TS scores of compounds. More specifically, each ion in the entire dataset was used sequentially as a target ion and the remaining ions in the dataset were ranked according to their pair-wise similarity to the selected target ion. Accordingly, the training sets consisted of subsets of the entire dataset and were obtained by selecting either the 8 or 10 most similar ions (for the anion and the cation dataset, respectively) to the target ion, or ions with TS values larger than a user-defined threshold. The training sets for the target anions were separately made up with ions on each column. Finally, the training sets were used to generate QSRR models for subsequent use for predicting the *a*- and *b*-values of their corresponding target ions, and ultimately the retention time. Each ion in the dataset served once as an external test compound for the external validation of its corresponding QSRR model.

5.2.4. QSRR modelling by GA-PLS

Genetic algorithm-partial least square (GA-PLS) regression was used for descriptor selection and QSRR model generation. Matlab (MathWorks, Natick, MA, USA) routines originally written by R. Leardi for GA-PLS [49] were modified to automatically run the algorithm 5 times for each ion in the dataset. Before running the algorithm almost constant descriptors, *i.e.*, descriptors having the same values for all but a few (maximum 5) compounds, were also eliminated from the training set of each ion. The GA was run using the following settings: the number of chromosomes in the original population was 50, the maximum selection probability was 20 variables/chromosome, average selection probability was 10 variables/chromosome, the mutation probability was 1%, the cross-over probability was 50%, and the number of evaluations prior to a backward elimination phase was 100 [12]. Subsequently, PLS regression was performed to relate the retention parameters (*a*- and *b*-values) to the selected

descriptors. For each repetition of the algorithm, the cross-validated R^2 (Q_{CV}^2) and the cross-validated root mean square error (*RMSECV*) for the training set were calculated. The Q_{CV}^2 and *RMSECV* were then used for the determination of the optimum number of LVs in PLS models. Leave-one-out approach was performed as an internal cross-validation method using the "venetian blind" approach for data sampling [12, 26]. Each model was then used to predict the *a*- and *b*-values of its corresponding target ion. The mean absolute error (*MAE*) of prediction for each ion was obtained by averaging 5 prediction errors obtained from 5 repetitions of the algorithm. The mean of *MAE* values were also obtained for each dataset by averaging all *MAE* values. In addition to the *MAE* values, the regression parameters, including slopes, intercepts, and coefficient of determination (R^2) values between the predicted and the measured *a*- and *b*-values, as well as retention time values, were obtained by carrying out linear regressions in MS Excel.

5.3. Results and Discussion

5.3.1. Role of structural similarity

Tanimoto similarity (TS) is based on 2-dimensional (2D) fingerprints which encode, as binary strings, the presence or absence of sub-structural fragments in a molecular structure. The TS is commonly used as a simple measure of inter-molecular structural similarity where a TS score is determined by a comparison of the strings. A TS score varies from zero, when no bit is in common between two structures, to unity when all bits are the same [50]. While there are many other structural similarity measures available, the TS score remains the most popular measure [50].

The role of structural similarity of ions to the unknown ion in QSRR modelling was investigated as follows. First, two types of training sets, consisting of either the most similar ions to the target ion or the least similar ions were compared in terms of the

prediction feasibility of the generated QSRR models. For this purpose, the ions in the database were ranked by their similarity to the target ion using the Tanimoto similarity index. Then for cations, the 10 most similar ions were chosen for one training set, and the 10 least similar ions for a second training set. (For anions, the eight most, and eight least similar ions formed the two training sets). By reducing the size of the training sets (to 10 and 8, from 20 in references [12, 36]), the ions included the training sets as a whole became more similar to the target ion, producing higher average TS scores for the training sets. This also facilitated the creation of training sets with a wide range of average TS scores (up to 0.80 for the cation dataset and 0.66 for anions, **Fig. 5.2**). **Fig. 5.1** shows the corresponding predicted *versus* measured *a*- and *b*-values for the studied anions and cations, where the results for anions were separately obtained on each column and then combined to give a pooled error for the method. From **Fig. 5.1-c**, **-d**, **-g** and **–h**, the models based on a training set of the least similar ions were not at all predictive, as is evident from the lack of correlation between predicted and measured values. On the other hand, when the training sets included the most similar ions **(Fig. 5.1-a, -b, -e and -f)**, the resultant QSRR models were predictive in that a much higher correlation between predicted and measured *a*- and *b*-values was observed. Worth noting are the strong correlations that were shown by the local models obtained from training sets which included only ions with an average TS score of at least 0.55 (red circles in **Fig. 5.1-a**, **-b**, **-e** and **-f**). The superior prediction accuracy of these points is evident from the very close proximity of these data points to the 45 degree line.

Next, the effect of similarity on the prediction accuracy in QSRR modelling was investigated using two measures: average TS scores of ions in the training set compared to the target ion (average TS), and the number of ions in the training set exceeding a threshold TS score based on the target ion (or the neighbour count [NC]) [34]. **Figure 5.2** shows *MAE* values plotted against the average TS scores of ions in the

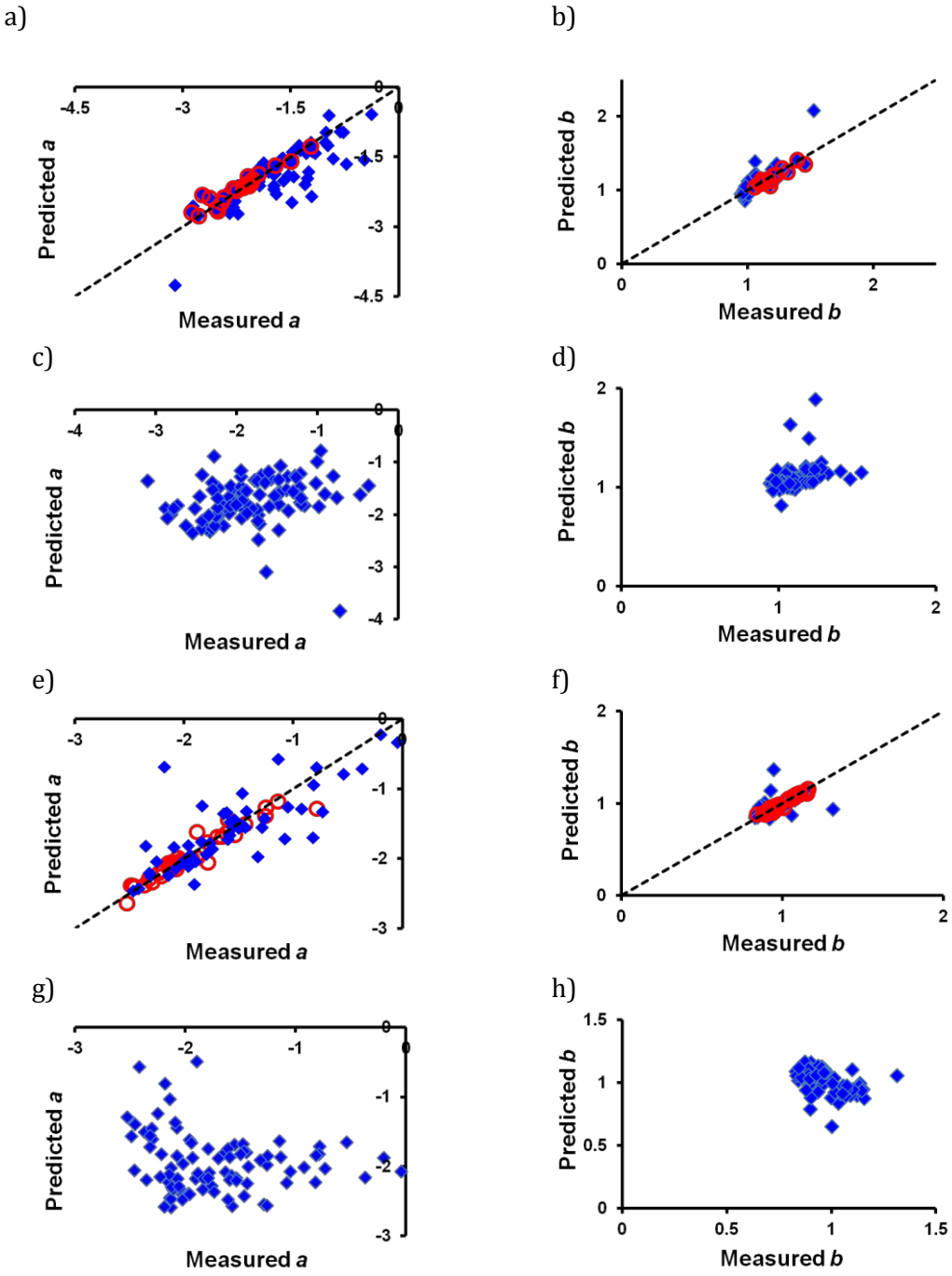


Figure 5.1 Correlations between predicted and measured *a*- and *b*-values for the combined anion datasets on three columns (AS20, AS19 and AS11HC) using the 8 most similar ions in the training sets to the unknown (target) ion (a and b) and the 8 least similar ions (c and d), and for the cation dataset on CS17 column using the 10 most similar ions (e and f) and the 10 least similar ions (g and h). The data with the average Tanimoto score of ≥0.55 for the ions in the training sets are indicated as the red circle markers.

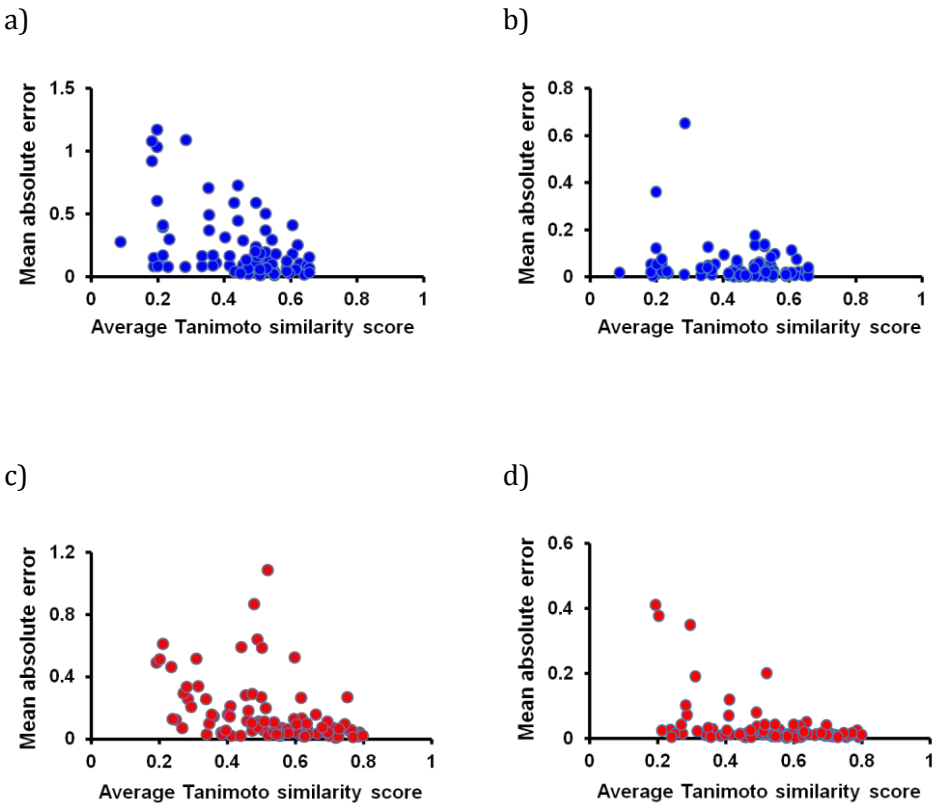
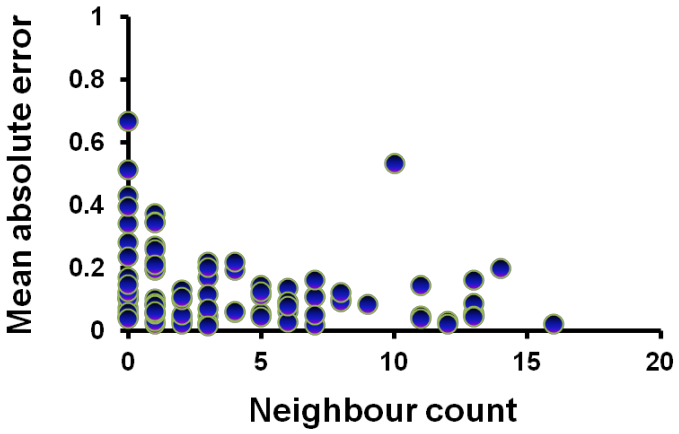


Figure 5.2 Mean absolute error (MAE) values *vs.* average Tanimoto score of ions in the training sets consisting of the 8 (a and b) and 10 (c and d) most similar ions to the unknown ion. The mean absolute error values were calculated for (a) the *a*-values and (b) the *b*-values of the anion compounds on the three columns (AS20, AS19 and AS11HC), and (c) the *a*-values and (d) the *b*-values of the cation compounds on the CS17 column.

training sets where the 10 and 8 most similar ions were included for the cation and anion datasets, respectively. Similar to **Fig. 5.1**, the results for anions were separately obtained on each column and then combined. The *MAE* values generally decreased with increasing average TS scores. Additionally, training sets having low TS scores produced variable prediction errors, ranging from low to high *MAE* values. These results imply that training sets of similar ions (*i.e.,* with higher average TS values) generate more predictive QSRR models, whereas the training sets of dissimilar ions create QSRR models with low or no predictability at all. This trend was more clearly evident for the larger cation dataset where the maximum value of average TS values in the training sets is around 0.8 (**Fig. 5.2-c** and **-d**). This result implies that even more accurate QSRR-models could be generated when a larger and more uniform dataset is available (so-called "Big Data QSRR-modelling" [12]), allowing the use of training sets comprised of highly similar ions (by the Tanimoto similarity index) to each unknown compound. **Fig. 5.1** also shows that the compounds that have training sets with higher average TS scores (represented by red circles) again give accurate predictions that are closer to the 45 degree line, whereas compounds with lower average TS scores in the training set (blue diamonds) fall further from the 45 degree line.

The cation dataset (*i.e.*, the larger dataset, containing a total of 87 ions) was utilised to explore the effect of the number of neighbours (NC) on the prediction accuracy. For this purpose, *MAE* values for models generated using the whole dataset were plotted against the NC, determined by applying a TS score of 0.6 as the limit value, as shown in **Fig. 5.3**. The results of the QSRR modelling for both *a*- (**Fig. 5.3-a**) and *b*-values (**Fig. 5.3-b**) showed that prediction errors generally decreased upon an increase in NC, except for one local model for *a*-values (in **Fig. 5.3-a**). This implies that the inclusion of more similar neighbours in the training set increases the prediction accuracy of the developed QSRR models. It is interesting to note that the errors for

a)



b)

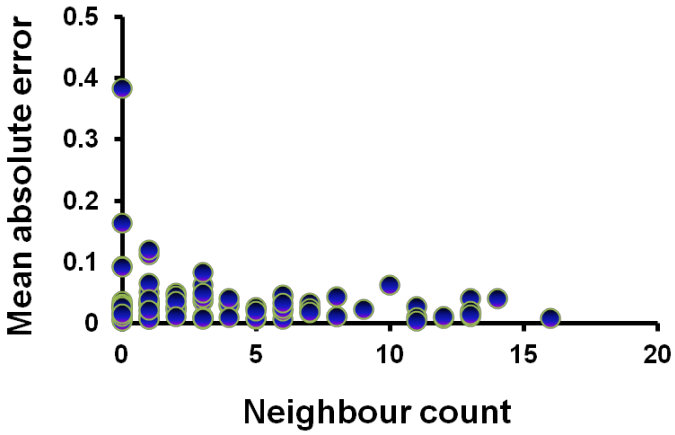


Figure 5.3 Mean absolute error (MAE) values *vs.* neighbour count in the training sets. The mean absolute error values were calculated for (a) the *a*-values and (b) the *b*-values for the entire cation dataset and the neighbour count were obtained as the number of neighbours having TS scores more than 0.6.

models created the NC method on the entire 87 cation dataset **(Fig. 5.3)** were generally somewhat smaller than those using the 10 most similar cations **(Fig. 5.2)**. This situation arises because the NC method produces larger training sets and this can lead to improved accuracy. It can therefore be seen that both the degree of similarity (*i.e.*, the average TS score or the TS cut-off used in the NC method) and the size of the training set should be maximized to give the most reliable predictions. These conclusions are similar to the results obtained for QSAR modelling [34].

5.3.2. Optimising and validating the QSRR model

Based on the findings from section 3.1, the approach for QSRR modelling adopted in the present study was to include in the training set for each target ion only neighbours having an adequate level of similarity defined by a TS threshold. By removing non-similar neighbours from the training set, the average similarity of the training set is increased, leading to an improvement in the prediction accuracy of the resultant models. Three TS cutoff values (0.4, 0.55, and 0.6) for the cation dataset and two (0.4 and 0.55) for the anion datasets were applied to generate the PLS models. Hence, 10 models were generated in total (five for *a*-values and five for *b*-values). The minimum number of compounds in the training sets was selected as seven to have sufficient compounds for modelling. The models generated with 7 ions resulted in similar prediction errors (compared to 10 ions) as well as valid Q_{CV}^2 values. **Table 5.3** summarises the characteristics of the developed PLS models. The optimal number of LVs was determined as less than or equal to three, and the number of descriptors ranged from 58 to 114. The Q_{CV}^2 values greater than 0.87 and low *RMSECV* values (<0.05) for all ten PLS models demonstrated the statistical significance and strong predictability of these models. Additionally, the predictive power of the proposed approach (*i.e.*, "federation of local QSRR models") was evaluated using the following parameters for external validation: (i) $MAE < 1$ (min) [12], (ii) the root mean square

Table 5.3 Predictive performance of QSRR models for *a*- and *b*-values. The Model name denotes cation or anion databases, *a*- or *b*-values, and the threshold value of Tanimoto score (TS). The *s* denotes the slope for the regression line (between predicted and measured *a*- and *b*-values) through the origin. No. Desc. is the number of descriptors and No. LV the number of latent variables utilised in the model. For all other definitions, refer to the text.

			Cross-validation			**External validation**			
Model	**No. Desc.**	**No. LV**	Q_{CV}^2	***RMSE CV***	***MAE***	***RMSEP***	$Q_{ext(F2)}^2$	***s***	$(R^2 - R_0^2)/ R^2$
Cation-a -TS-0.4*	91	2	0.97	0.04	0.13	0.22	0.72	1.00	0.18
Cation-b -TS-0.4	58	2	0.88	0.01	0.03	0.03	0.88	1.00	0.02
Cation-a -TS-0.55	111	2	0.94	0.03	0.09	0.15	0.85	0.99	0.05
Cation-b -TS-0.55	81	2	0.95	0.005	0.014	0.02	0.93	1.00	0.005

Cation-a -TS-0.6	83	2	0.88	0.03	0.06	0.08	0.81	1.00	0.001
Cation-b -TS-0.6	80	2	0.95	0.003	0.011	0.01	0.96	1.00	0.002
Anion-a -TS-0.4*	114	2	0.98	0.04	0.12	0.18	0.66	0.98	0.06
Anion-b -TS-0.4	104	2	0.97	0.01	0.03	0.05	0.71	0.99	0.07
Anion-a -TS-0.55*	74	3	0.99	0.03	0.11	0.14	0.59	1.00	0.05
Anion-b -TS-0.55*	114	3	0.98	0.01	0.04	0.05	0.68	0.99	0.38

* Models failed to meet all the criteria.

error of prediction (*RMSEP*) [51, 52], (iii) the predictive squared correlation coefficient $Q_{ext(F2)}^2$ >0.7 [53], (iv) 0.85 < s < 1.15, and (v) $(R^2 - R_0^2)/R^2$ <0.1 [37], where R^2 is the coefficient of determination between predicted and measured values, and s and R_0^2 are respectively, the slopes and the coefficient of determination for regression lines through the origin. The $Q_{ext(F2)}^2$ and *RMSEP* were calculated by [51]:

$$Q_{ext(F2)}^2 = 1 - \frac{\sum_i^m (y_{i,t} - \hat{y}_{i,t})^2}{\sum_i^m (y_{i,t} - \overline{y_t})^2} \quad \textbf{(5.2)}$$

$$RMSEP = \left[\frac{1}{m}\sum_{i=1}^m (y_{i,t} - \hat{y}_{i,t})^2\right]^{0.5} \quad \textbf{(5.3)}$$

where $y_{i,t}$ are the measured values for external test ions, $\hat{y}_{i,t}$ are the predicted values for external test ions, $\overline{y_t}$ is the mean measured value for external test ions, m is the total number of external test ions. All parameters in **Table 5.3** were obtained by averaging the values from the PLS models developed for each compound. Six out of the ten generated PLS models satisfied all of the criteria mentioned above, with the remaining four models (PLS model for a-values of cations based on the TS cutoff of 0.4 [$(R^2 - R_0^2)/R^2$ of 0.18], a-values of anions with a TS cutoff of 0.4 ($Q_{ext(F2)}^2$ of 0.66), and a- and b-values of anions with a TS cutoff of 0.55 [$(R^2 - R_0^2)/R^2$ of 0.38 (b-values), and $Q_{ext(F2)}^2$ of 0.59 and 0.68, respectively), thereby failing to meet the validation criteria mentioned above. Additionally, the *MAE* values for the PLS models using cation and anion data generally decreased as the TS cutoff values increased. This implied that more predictive and accurate PLS models can be generated when the training sets have a higher level of similarity with the target ion.

Furthermore, the ten generated models were compared and evaluated using the sum of ranking difference (SRD) approach [54-56] and a free-to-access tool provided in reference [55]. The input matrix consists of eight rows and eleven columns, representing 8 validation criteria and both 10 models and reference values in **Table 5.3**. Row-minimum conditions for parameters (*No. Desc.*, *RMSECV*, *MAE*, *RMSEP*, and (R^2

$- R_0^2)/R^2$) and row-maximum ones (Q_{CV}^2, $Q_{ext(F2)}^2$, and s) were used as reference values which were added in the last column of the input matrix. **Fig. 5.4** shows the SRD values for all the ten models are smaller (<20%) than those for models produced randomly (*i.e.*, simulated models) shown as theoretical distribution of the random SRD values, which implies that the generated models are valuable. In addition, models for *a*- and *b*-values of cations based on a TS cutoff of 0.6 were identified as the best models, showing the smallest SRD values, which is consistent with the discussion above. Additionally, these two models represent the same quality, as found indistinguishable in **Fig. 5.4** and by showing the same ranking of 2 (**Table 5.4**). The ranking of models based on the SRD approach was generally in accord with the magnitude of TS cutoff values, as expected.

Figure 5.5 illustrates the correlations between predicted and measured values from all the PLS models developed using the cation data for the three TS cutoff values. The fit of the data onto the 45 degree line indicates the predictive ability of the generated PLS models since each data point represents the corresponding local PLS model. As expected, the strongest correlation was obtained for the training set with the highest TS cutoff value of 0.6 **[Figs. 5.4(e)** and **5.4(f)]**, resulting in PLS models with the highest predictive power (corresponding to the lowest *MAE* values). Similar results were obtained for the anion data **(Fig. 5.6)** although there are fewer data points in the graphs due to the small size of the data sets.

5.3.3. Application of QSRR models for *a* and *b* values to the prediction of retention times

The predicted *a*- and *b*-values of the QSRR models were fitted into the LSS model equation **(Eq. 5.1)** to check the applicability of the models for retention time prediction. For this purpose, the *a*- and *b*-values based on the highest TS cutoff values (0.55 and 0.6 for the cation data) were substituted into **Eq. 5.1** at different eluent concentrations (5, 10, 20, 30, and 40 mM MSA eluent on CS17 column). The resultant retention time

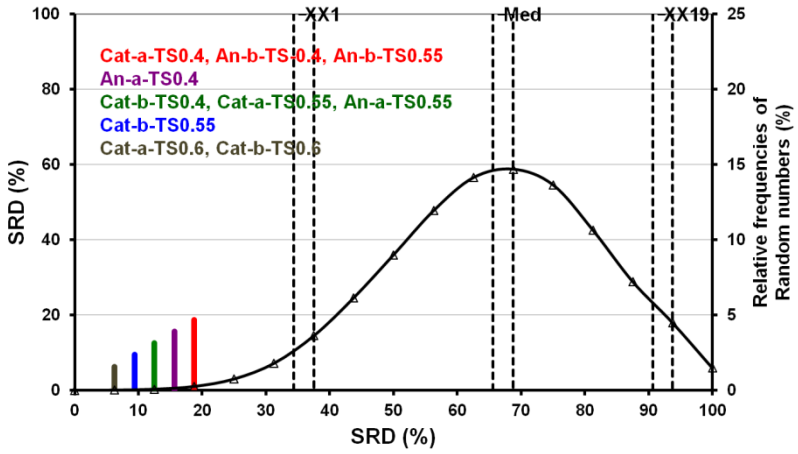


Figure 5.4 Comparison of the models using sum of ranking differences (SRD) with ties (repeated observations). Scaled SRD values (between 0 and 100) are plotted on both the *x* axis and left *y* axis, and the right *y* axis shows the relative frequencies (black curve). Probability levels (XX1 = 5% limit, Med = median, and XX19 = 95% limit) are also given.

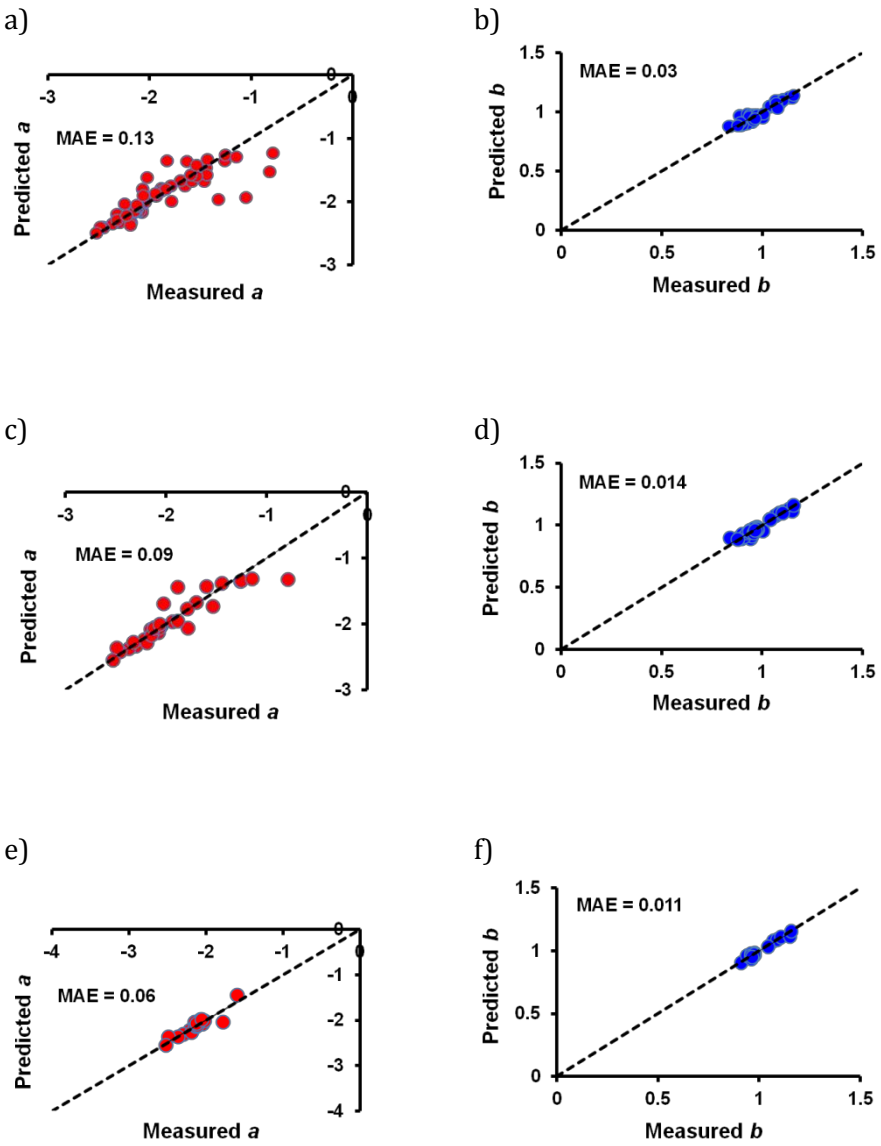


Figure 5.5 Correlations between predicted and measured values for the cation dataset. The plots present the data for the QSRR models for both *a*-and *b*-values, generated using only similar neighbours in the training sets based on the Tanimoto score cutoff of 0.4 [(a) and (b)], 0.55 [(c) and (d)], and 0.6 [(e) and (f)].

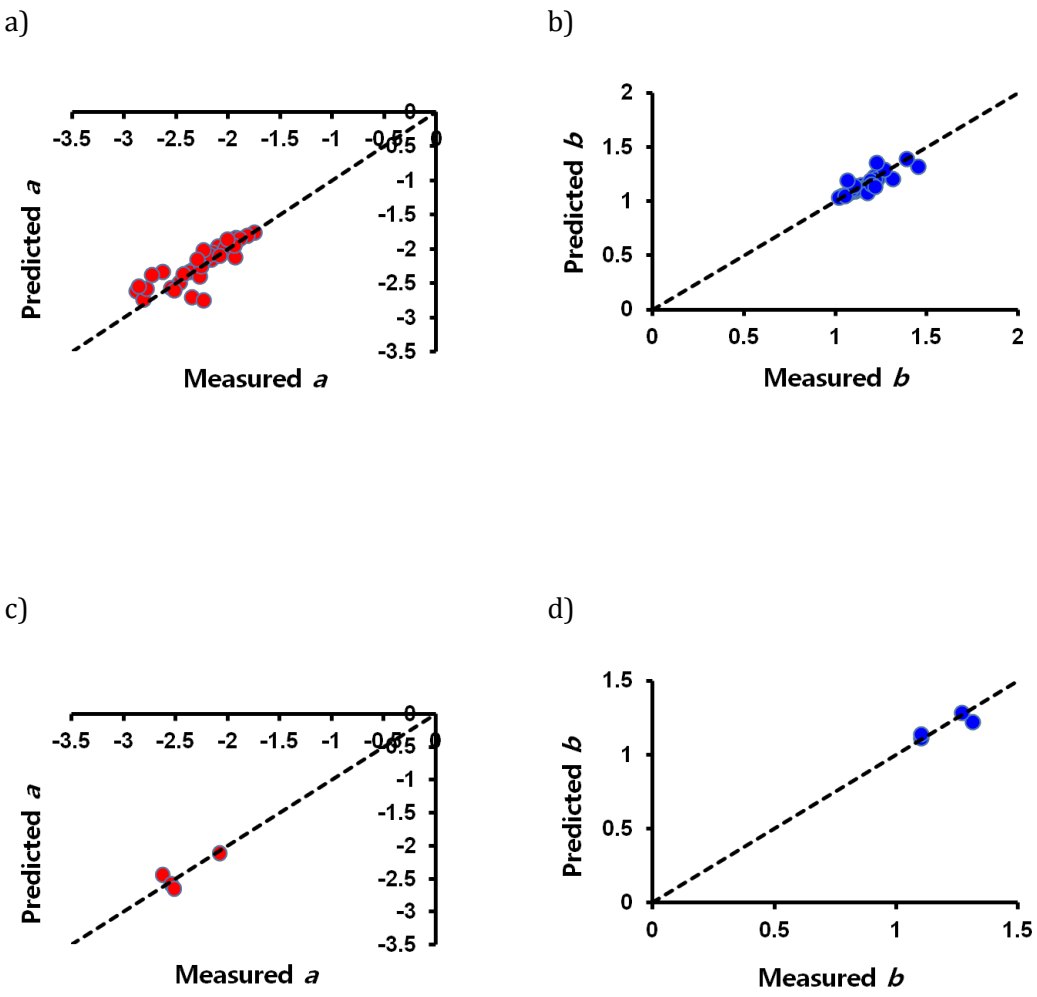


Figure 5.6 Correlations between predicted and measured values for the anion datasets. The plots present the data for the QSRR models for both *a*-and *b*-values, generated using only similar neighbours in the training sets based on the TS cutoff of 0.4 [(a) and (b)] and 0.55 [(c) and (d)].

Table 5.4 Results of sum of ranking differences (SRD) analysis.

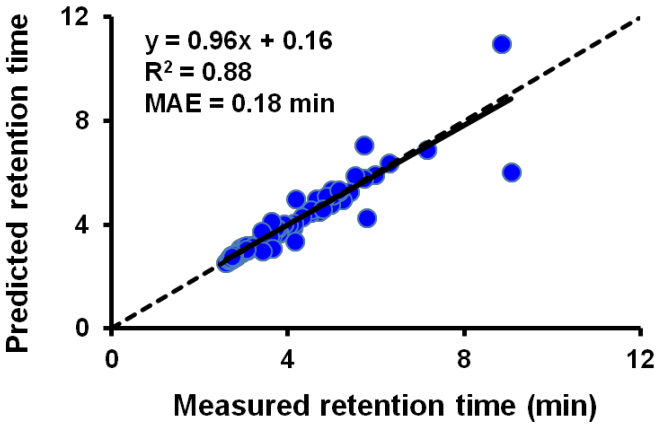
Model	SRD	p%	
		x < SRD > =x	
Cation-a-TS-0.6	2	3.0E-03	1.6E-02
Cation-b-TS-0.6	2	3.0E-03	1.6E-02
Cation-b-TS-0.55	3	1.6E-02	2.3E-02
Cation-b-TS-0.4	4	2.3E-02	8.8E-02
Cation-a-TS-0.55	4	2.3E-02	8.8E-02
Anion-a-TS-0.55	4	2.E-02	8.8E-02
Anion-a-TS-0.4	5	8.8E-02	0.11
Cation-a-TS-0.4	6	0.11	0.35
Anion-b-TS-0.4	6	0.11	0.35
Anion-b-TS-0.55	6	0.11	0.35
XX1	12	3.3	6.5
Q1	18	23	34
Med	22	49	62
Q3	24	64	76
XX19	30	94	98

values were then plotted against the corresponding measured values. **Fig. 5.7** shows the correlation plots of predicted *versus* measured retention times and the corresponding regression and external validation parameters are presented in **Table 5.5**. Good correlation was shown for retention time prediction for the cation data based on the TS cutoff value of 0.6, with the low prediction error (*i.e.*, *RMSEP* of 0.44 min) which includes the error of 0.04 min in retention times obtained by using the *a* and *b* values for the LSS model. On the other hand, poor correlation was found for the data with the TS cutoff value of 0.55. Indeed, the external validation for the corresponding model was not successful for three criteria (*s* of 0.80, $(R^2 - R_0^2)/R^2$ of 0.38, and $Q_{ext(F2)}^2$ of 0.66). Additionally, the QSRR models for the anion dataset using the TS cutoff value of 0.55 for predicting the *a*- and *b*-values were found inappropriate for further use in retention time prediction, given that these models fell outside the criteria for external validation (Section 3.2). Thus, it can be concluded that only QSRR models for the *a*- and *b*-values based on TS cutoff value of 0.6 (for cation data) are sufficiently accurate for further use in retention time prediction, resulting in an average *MAE* value of 0.2 min for the retention time predictions.

5.4. Conclusions

In the present study, the structural similarity between ions in the training sets and the target (unknown) ion has been found to be a significant factor in generating predictive QSRR models in ion chromatography. The approach of "federation of local models" using similar neighbours (identified by employing a Tanimoto similarity (TS) score threshold), to the target ion was employed successfully to construct QSRR models for *a*- and *b*-values in the isocratic LSS model in ion chromatography. The results of external validation showed that TS score cutoff values $\geq$0.6 are needed to develop accurate and predictive QSRR models for *a*- and *b*-values ($Q_{ext(F2)}^2$ >0.8, and Mean absolute error

a)



b)

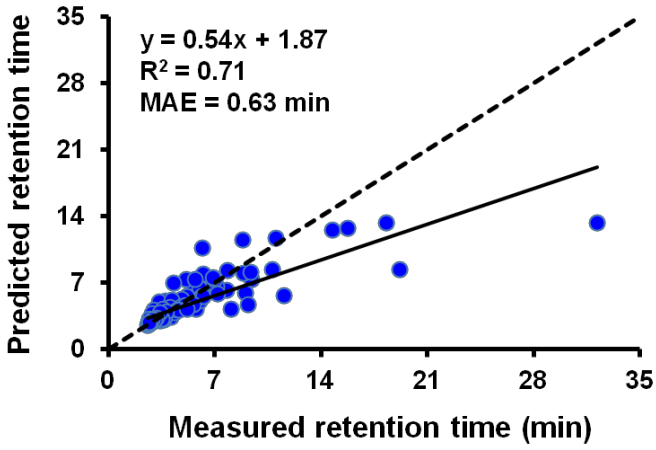


Figure 5.7 Correlations between predicted and measured retention times using predicted *a*- and *b*-values of QSRR models, generated using only similar neighbours in the training sets based on the TS cutoff of 0.6 (a) and 0.55 (b) using the cation dataset. The dashed and full lines are the 45 degree and regression lines, respectively.

Table 5.5 Predictive performance of QSRR models for t_R-values. The *s* denotes the slope for regression line (between predicted and measured t_R-values) through the origin.

	MAE **(min)**	***RMSEP*** **(min)**	$Q_{ext(F2)}^2$	s_o	$(R^2 - R_0^2)/ R^2$
Cation-t_R-TS-0.6	0.18	0.44	0.87	1.00	0.002
Cation-t_R-TS-0.55	0.63	0.91	0.66	0.80	0.38

values <0.1). This in turn allows accurate retention time predictions (Mean Absolute Error of 0.2 min) for unknown ions under a wide range of eluent concentrations. Furthermore, larger datasets (cations *vs.* anions) have an advantage in that they allow the construction of training sets with a higher degree of similarity, *i.e.*, a higher average TS score and neighbour count based on a higher TS score cut off. This implies that the establishment of larger and more homogeneous datasets (*i.e.*, datasets comprised of clusters of ions that are of high structural similarity) is needed for the successful development of larger numbers of local models, where the training sets have sufficient similar neighbours to the target ions.

The proposed approach allows retention time predictions with excellent accuracy for unknown ions under a broad range of eluent conditions on various columns, enabling rapid scoping method development in ion chromatography. Future work includes investigating alternative similarity searching strategies within the "federation of local QSRR models" concept, which can increase the number of candidate ions being eligible for local modelling.

5.5. References

[1] B.K. Ng, T.T.Y. Tan, R.A. Shellie, G.W. Dicinoski, P.R. Haddad, Computer-assisted simulation and optimisation of retention in ion chromatography, TrAC 80 (2016) 625-635.

[2] E. Tyteca, S.H. Park, R.A. Shellie, P.R. Haddad, G. Desmet, Computer-assisted multi-segment gradient optimization in ion chromatography, J. Chromatogr. A 1381 (2015) 101-109.

[3] V. Drgan, D. Kotnik, M. Novic, Optimization of gradient profiles in ion-exchange chromatography using computer simulation programs, Anal. Chim. Acta 705 (2011) 315-321.

[4] V. Drgan, M. Novic, M. Novic, Computational method for modeling of gradient separation in ion-exchange chromatography, J. Chromatogr. A 1216 (2009) 6502-6510.

[5] A.D. Sosimenko and P.R. Haddad, Computer optimization in IC-ii a systematic evaluation of linear retention models for anions, J.Chromatogr. 546 (1991) 37-59.

[6] J. Madden, P.R. Haddad, Retention models in ion chromatography and their use in computer optimization of eluent composition, TrAC 15 (1996) 531-537.

[7] R.D. Rocklin, C.A. Pohl, J.A. Schibler, Gradient elution in ion chromatography, J. Chromatogr. A 411 (1987) 107-119.

[8] P. Jandera, J. Churacek, Gradient elution in liquid chromatography I. The influence of the composition of the mobile phase on the capacity ratio (retention volume, band width, and resolution) in isocratic elution-theoretical considerations, J.Chromatogr. 91 (1974) 207-221.

[9] S.H. Park, R.A. Shellie, G.W. Dicinoski, G. Schuster, M. Talebi, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl, Enhanced methodology for porting ion chromatography retention data, J. Chromatogr. A 1436 (2016) 59-63.

[10] R.A. Shellie, B.K. Ng, G.W. Dicinoski, S.D. Poynter, J.W. O'Reilly, C.A. Pohl, P.R. Haddad, Prediction of analyte retention for ion chromatography separations performed using elution profiles comprising multiple isocratic and gradient steps, Anal. Chem. 80 (2008) 2474-2482.

[11] P.R. Haddad, P.E. Jackson, Ion chromatography: principles and applications, in: Journal of Chromatography Library, Elsevier, Amsterdam, The Netherlands, 1990.

[12] E. Tyteca, M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad, Towards a chromatographic similarity index to establish localized Quantitative Structure-Retention Models for retention prediction: use of retention factor ratio, J. Chromatogr. A 1486 (2017) 50-58.

[13] R. Kaliszan, QSRR: Quantitative Structure-(Chromatographic) Retention Relationships, Chem. Rev. 107 (2007) 3212-3246.

[14] G. Carlucci, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure-retention relationships, Anal. Chim. Acta 601 (2007) 68-76.

[15] Š. Ukić, M. Novak, P. Žuvela, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part I: Conventional QSRR approach, Chromatographia 77 (2014) 985-996.

[16] Š. Ukić, M. Novak, A. Vlahović, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part II: Artificial intelligence QSRR approach, Chromatographia 77 (2014) 997-1007.

[17] T. Baczek, R. Kaliszan, Combination of linear solvent strength model and quantitative structure–retention relationships as a comprehensive procedure of approximate prediction of retention in gradient liquid chromatography, J. Chromatogr. A 962 (2002) 41–55.

[18] E. Daghir-Wojtkowiak, S. Studzińska, B. Buszewski, R. Kaliszan, M.J. Markuszewski, Quantitative structure–retention relationships of ionic liquid cations in characterization of stationary phases for HPLC, Anal. Methods 6 (2014) 1189-1196.

[19] R. Kaliszan, M.A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure–retention relationships, J. Chromatogr. A 855 (1999) 455–486.

[20] A. Plenis, L. Konieczna, N. Miekus, T. Baczek, Development of the HPLC method for simultaneous determination of lidocaine hydrochloride and tribenoside along with

their impurities supported by the QSRR approach, Chromatographia 76 (2013) 255-265.

[21] N. Kritikos, A. Tsantili-Kakoulidou, Y.L. Loukas, Y. Dotsikas, Liquid chromatography coupled to quadrupole-time of flight tandem mass spectrometry based quantitative structure-retention relationships of amino acid analogues derivatized via n-propyl chloroformate mediated reaction, J. Chromatogr. A 1403 (2015) 70-80.

[22] P.E. Morgan, D.J. Barlow, M. Hanna-Brown, R.J. Flanagan, Artificial neural network modelling of the retention of acidic analytes in strong anion-exchange HPLC: Elucidation of structure-retention relationships, Chromatographia 75 (2012) 693-700.

[23] K. Gorynski, B. Bojko, A. Nowaczyk, A. Bucinski, J. Pawliszyn, R. Kaliszan, Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds, Anal. Chim. Acta 797 (2013) 13-19.

[24] M. Goodarzi, R. Jensen, Y. Vander Heyden, QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions, J. Chromatogr. B 910 (2012) 84-94.

[25] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure–retention relationship studies, J. Chromatogr. A 988 (2003) 261-276.

[26] M. Talebi, G. Schuster, R.A. Shellie, R. Szucs, P.R. Haddad, Performance comparison of partial least squares-related variable selection methods for quantitative

structure retention relationships modelling of retention times in reversed-phase liquid chromatography, J. Chromatogr. A 1424 (2015) 69-76.

[27] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemom. Intell. Lab. Syst. 76 (2005) 185-196.

[28] P. Zuvela, J.J. Liu, K. Macur, T. Baczek, Molecular descriptor subset selection in theoretical peptide quantitative structure-retention relationship model development using nature-inspired optimization algorithms, Anal. Chem. 87 (2015) 9876-9883.

[29] V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, Prediction of capillary gas chromatographic retention times of fatty acid methyl esters in human blood using MLR, PLS and back-propagation artificial neural networks, Talanta 83 (2011) 1014-1022.

[30] J. Ghasemi, S. Saaidpour, QSRR Prediction of the Chromatographic Retention Behavior of Painkiller Drugs, J. Chromatogr. Sci. 47 (2009) 156-163.

[31] C.B. Mazza, C.E. Whitehead, C.M. Brenernan, S.M. Crarner, Predictive quantitative structure-retention relationship models for ion-exchange chromatography, Chromatographia 56 (2002) 147-152.

[32] M. Song, C.M. Breneman, J. Bi, N. Sukumar, K.P. Bennett, S. Cramer, N. Tugcu, Prediction of protein retention times in anion-exchange chromatography systems using support vector regression, J. Chem. Inf. Comput. Sci. 42 (2002) 1347-1357.

[33] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions, J. Mol. Graph. Model. 23 (2005) 503-523.

[34] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, S.K. Kearsley, Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR, J. Chem.

Inf. Comput. Sci. 44 (2004) 1912-1928.

[35] M.A. Johnson, G.M. Maggiora, Concepts and applications of molecular similarity, John Wiley & Sons, New York, 1990.

[36] M. Talebi, S.H. Park, M. Taraji, Y. Wen, R.I.J. Amos, P.R. Haddad, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan, Retention time prediction based on molecular structure in pharmaceutical method development: A perspective, LCGC North America 34 (2016) 550-558.

[37] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: Validation in the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69-77.

[38] P. Zakaria, G.W. Dicinoski, B.K. Ng, R.A. Shellie, M. Hanna-Brown, P.R. Haddad, Application of retention modelling to the simulation of separation of organic anions in suppressed ion chromatography, J. Chromatogr. A 1216 (2009) 6600-6610.

[39] MarvinSketch, ChemAxon 2016, chemaxon.com [Accessed: April 2016].

[40] T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J. Comp. Chem. 17 (1996) 490-519.

[41] T.A. Halgren, Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions, J. Comp. Chem. 17 (1996) 520-552.

[42] T.A. Halgren, Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94, J. Comp. Chem. 17 (1996) 553-586.

[43] T.A. Halgren, R.B. Nachbar, Merck molecular force field. IV. Conformational energies and geometries for MMFF94, J. Comp. Chem. 17 (1996) 587-615.

[44] M.J. Vainio, M.S. Johnson, Generating conformer ensembles using a multiobjective genetic algorithm, J. Chem. Inf. Model. 47 (2007) 2462-2474.

[45] J.S. Puranen, M.J. Vainio, M.S. Johnson, Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery, J. Comp. Chem. 31 (2010) 1722-1732.

[46] J.J. Stewart, Optimization of parameters for semiempirical methods VI: more modification to the NDDO approximations and re-optimization of parameters, J. Mol. Model. 19 (2013) 1-32.

[47] MOPAC 2012, Stewart Computational Chemistry, Colorado Springs: CO, USA, OpenMOPAC.net.

[48] Dragon 6.0, Talete, Milano, Italy, 2014, talete.mi.it [Accessed: April 2016].

[49] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.

[50] P. Willett, Similarity-based virtual screening using 2D fingerprints, Drug Discovery Today 11 (2006) 1046-1053.

[51] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, J. Chemometr. 24 (2010) 194-201.

[52] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, Chemom. Intell. Lab. Syst. 152 (2016) 18-33.

[53] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection, J. Chem. Inf. Model. 52 (2012) 2044-2058.

[54] K. Héberger, Sum of ranking differences compares methods or models fairly, TrAC 29 (2010) 101-109.

[55] K. Kollár-Hunek, K. Héberger, Method and model comparison by sum of ranking differences in cases of repeated observations (ties), Chemom. Intell. Lab. Syst. 127

(2013) 139-146.

[56] A. Racz, D. Bajusz, K. Heberger, Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters, SAR and QSAR in Environmental Research 26 (2015) 683-700.

Chapter 6

Chromatographic similarity-based QSRR modelling

6.1. Introduction

Quantitative Structure-Retention Relationships (QSRR) are statistically-derived mathematical relationships between chromatographic retention parameters and molecular descriptors encoding the chemical structure of analytes [1]. QSRRs have been extensively investigated in topics such as the classification and characterisation of stationary phases [1-3], the elucidation of retention mechanisms [1, 4, 5], identification of the most significant descriptors relating to chromatographic retention [6-9], and the retention prediction of unknown analytes [10-14]. The QSRR approach used to predict retention times of unknown analytes can speed up the "scoping" phase of chromatographic method development in that the selection of broad chromatographic conditions (such as stationary and mobile phase composition) can be simplified by retention models based only on the structures of new analytes, resulting in the minimisation of experimentation [15, 16]. This goal can be achieved only when the QSRR models have an appropriate level of prediction accuracy.

Prediction errors in modelling may arise either due to the use of poorly chosen (or insufficient) compound properties in the model generation or the inclusion of diverse classes of compounds for a single QSRR model [17]. Accordingly, one strategy to improve model accuracy can be to introduce a compound classification (or clustering) procedure prior to QSRR modelling, in which compounds having similar characteristics to the target ion are included in the same subclass of the entire dataset [15, 17]. Here, the term “target ion” refers to the analyte ion for which retention is to be predicted. Subsequently, the most favourable subclass for a given target compound is used to generate the corresponding QSRR model [15, 17-19]. Compound classification (or

clustering) has been performed using various similarity measures, which include the compound interactions with both the stationary phase and the mobile-phase [17], log*D* profile (according to pH) [18], Tanimoto similarity index such as in Chapter 5 [15, 19, 20], and a combination of log*D* and retention factor [15].

Tyteca *et al*. [15] have proposed the retention factor ratio (or *k*-ratio) as a chromatographic similarity index, where clustered compounds based on a selected *k*-ratio threshold value are used for modelling. In other words, to create a cluster that could be used as a training set to form a model for the target compound, only compounds having absolute *k*-ratio values lower than a threshold value were included in the training set to generate a partial least squares (PLS) model. This approach resulted in low prediction errors (mean absolute error (*MAE*) <1 min) for the three liquid chromatography modes (HILIC, RPLC and IC), which is a sufficient level of accuracy for "scoping" method development. However, the *k*-ratio-based-clustering approach has a limitation in that when a new target compound is used it is not possible to identify the training set with most similar retention times due to the lack of *a priori* retention information for the target compound. Thus, a dual filter approach, combining the structural Tanimoto similarity (TS)-based-filter with the *k*-ratio filter, was also investigated using the HILIC and RPLC datasets [15]. Unfortunately, the employed dual filter approach did not improve predictions in HILIC and RPLC, presumably because the TS filter (used as the first filter) did not adequately reflect the chromatographic similarity of the compounds, possibly due to the complex mechanisms at play in HILIC and RPLC [15].

In the case of IC, the dataset used in the study mentioned above (which included 49 inorganic and low-molecular weight organic anions) was not large enough to apply the dual filter approach. In Chapter 5, compound clustering based on a TS threshold value larger than 0.6 resulted in predictive QSRR models for inorganic and low-

molecular weight organic anions as well as some larger organic cations, having low prediction errors (*MAE* <1 min) [20]. This is presumably due to the relatively simple retention mechanism in IC, mainly dominated by the charge density (ratio of charge and solvated size) of ions, and the simple chemical structures of the ions, which enabled the Tanimoto similarity index to reflect adequately the retention in IC [15]. In Chapter 4, as an alternative to directly predicting the retention times of ions, two retention parameters (*a* and *b*) in the linear solvent strength (LSS) model in IC (log*k* = *a* − *b*log[eluent]) have been successfully modelled using QSRR for the prediction of retention times of low molecular weight anions on three Thermo Fisher columns (AS20, A19 and AS11HC) [20]. The QSRR modelling for the prediction of these two retention parameters (*a*- and *b*-values) is an attractive approach in that the predicted *a*- and *b*-values in the LSS model can provide accurate retention time predictions over a broad range of eluent concentrations under isocratic and gradient conditions, as well as multi-step elution profiles comprising sequential isocratic and gradient steps [21]. The LSS model typically applies for inorganic and small ions, for which the retention mainly results from electrostatic interactions. Although larger organic ions are often separated based on mixed-mode retention comprised of both ion-exchange and hydrophobic interactions, the LSS model can be applicable to predict the retention of these organic ions when their hydrophobic interaction is sufficiently reduced by adding organic modifier to the eluent [21].

Two physicochemical descriptors, log*P* and log*D,* are often used to describe the hydrophobicity of molecules as the molecular and ionic forms, respectively. The logarithm of the octanol-water partition coefficient (log*P*) has been frequently employed as a descriptor to represent the hydrophobic interaction in RPLC [1, 3, 22-24]. In these QSRR studies, direct linear relationships between log*P* and the logarithm of the retention factor extrapolated to pure water (log k_w) have been successfully

derived:

$$\log k_w = a + b\log P \quad \textbf{(6.1)}$$

where the $\log k_w$ is a standardised retention parameter determined as the intercept of the linear solvent strength model in RPLC (or the extrapolated retention for pure water mobile phase) [1]. The QSRR model **(Eq. 6.1)** can be used to predict the retention for neutral compounds and congeneric ionic compounds [22], and has also been applied to IC for ionic liquids, confirming their hydrophobic interactions [25]. For ionisable compounds, the logarithm of the octanol-water distribution coefficient ($\log D$), rather than $\log P$, is generally used to elucidate hydrophobicity [26]. As an example, the retention prediction of ionisable compounds on a C18 column under gradient conditions has been improved by using compounds having similar $\log D$ profile (over a specified pH range) to create the models [18].

In the present chapter, chromatographic similarity-based localised QSRR modelling for *a*- and *b*-values in the LSS model in ion chromatography was investigated using larger organic cations of pharmaceutical interest, to generate sufficiently predictive and accurate models available for rapid scoping method development. Several filtering approaches to cluster only the ions similar to each target ion into the training set were compared based on the errors in the prediction of retention times, such as *MAE* and root mean squared error of prediction (*RMSEP*), where predicted retention times were calculated by fitting predicted *a*- and *b*-values of models to the LSS model. The *k*-ratio filtering based on a cut-off value of 1.2 was initially investigated. Next, a dual filter, combining the *k*-ratio filter with the Tanimoto similarity (TS) filter, was studied to address the limitation in the *k*-ratio filtering, which is that it is impractical to use for retention prediction for new target ions. In the dual filter, the most similar ion to the target ion was selected by its TS score (the first filter) and then a *k*-ratio filter (the second filter) was applied using the retention factor of the most

similar ion selected in the first filter as the reference [15]. Finally, the primary filter used earlier was modified to include a TS filter followed by either $\Delta\log P$- or $\Delta\log D$-indices. The purpose of such a combined primary filter was to select the most chromatographically similar ion to the target ion prior to the *k*-ratio filtering. It was hoped that the combined filtering approach could adequately encompass all the interactions associated with the IC retention of larger organic cations and the stationary phase, including the hydrophobic interactions which can be addressed by the introduction of the $\Delta\log P$ index, with the aim of providing clusters of ions with more "chromatographic similarity". The performance of the developed QSRR models for *a*-, *b*- and t_R- values was characterised by evaluating external validation parameters such as $Q_{ext(F2)}^2$ (cross-validated R^2), *MAE* and *RMSEP*.

6.2. Materials and Methods

6.2.1. Datasets

The dataset consisted of the *a*- and *b*- values for 87 organic monovalent cations (molecular mass up to 506) on a Thermo Fisher IonPac CS17 (2 mm i.d. × 250 mm) column **(Table 5.2)**. The *a*- and *b*-values were estimated from the LSS model using retention time data collected under five isocratic eluent compositions (5, 10, 20, 30, and 40 mM MSA with 36% ACN). Retention data for these ions were collected using a Dionex (Sunnyvale, CA, USA) ICS-3000 Ion Chromatography system, comprising a dual gradient pump unit (DP), a dual eluent generator unit (EG), dual suppressed conductivity detector compartment (DC), variable wavelength UV detector (VWD) and autosampler (AS). All analyses were performed on a CS17 analytical column (2 mm i.d. × 250 mm) with its associated guard column (2 mm i.d. × 50 mm) at a column temperature of 30°C. The injection volume was 10 μL and the eluent flow-rate was 0.25 mL/min. The methanesulfonic acid (MSA) was pumped at 0.16 mL/min and mixed with

acetonitrile (ACN) at 0.09 mL/min through a T-piece connector followed by a gradient mixer (Dionex GM-4 2mm). Milli-Q water (18.2 MΩ; Merck-Millipore, Bayswater, Australia) at 0.5 mL/min was supplied to the EG, continuously regenerated trap column (CR-TC) and degasser using an additional pump (Jasco PU-2089i Plus, Tokyo, Japan). 58 Cations were detected by UV detection at 220 nm and two cations (synephrine and tyramin) were at 270 nm (**Table 5.2**). The remaining 29 cations were analysed by suppressed conductivity detection at 35°C, using a CSRS® 300 (2 mm) electrolytic suppressor (**Table 5.2**). Chromeleon software (version 6.80) was used for instrument control and data acquisition.

6.2.2. Molecular descriptors

Molecular descriptors were calculated by the method employed in Chapter 5. Briefly, 2D structures of ions were drawn using MarvinSketch version 6.2.1 (ChemAxon, Budapest, Hungary) [27]. Initial conformational searches were performed using the Merck Molecular Force Field (MMFF94) [28-31] (in Balloon [32, 33]) to find the 50 lowest energy 3D-conformers. Geometrical optimisation for the lowest energy conformers were then carried out in water by the semi-empirical Parametric Method 7 (PM7) [34], implemented in Molecular Orbital PACkage (MOPAC) [35]. Finally, molecular descriptors were calculated by the Dragon 6.0 software (Talete, Milano, Italy) [36] after uploading the geometrically optimised 3D structures. The 4885 descriptors calculated initially were reduced to 570 descriptors by automatic screening, removing descriptors with constant values, with at least one missing value, with standard deviation ≤ 0.0001, and with an absolute pair-wise correlation $\geq$0.9. The resulting 570 descriptors were used for genetic algorithm combined with partial least squares regression (GA-PLS) modelling.

6.2.3. Similarity searching for training sets

A training set to generate each local QSRR model was obtained by including similar ions to each target ion based on various similarity filters. **Figure 6.1** outlines the similarity searching by three types of filters (see below). For the *k*-ratio filter-based similarity searching which follows the method in reference [15], the ions in the dataset were ranked according to the absolute values of the retention factor ratio (or *k*-ratio) of ions to the target ion and those ions having an absolute value of *k*-ratio <1.2 were included in the training sets for the corresponding target ion (*k*-ratio filter, **Fig. 6.1a**). In an analogous process, similarity searching by log*P*-ratio and log*D*-ratio filters was also performed for completeness. The threshold values for these two methods were 1.4 and 1.2, respectively.

For the dual filter-based similarity searching, the most similar ion to the target ion was selected as a surrogate of the target ion by the first filter (see below) and ions in the database that were most similar to the surrogate ion were then clustered into the training set using the *k*-ratio (<1.2) filter as the second filter. Two approaches were investigated to choose the surrogate ion at the first filter: The first was to choose the ion having the highest pairwise Tanimoto similarity (TS) score to the target ion (Dual filter, **Fig. 6.1b**) and the second approach was to select the surrogate ion by the TS filter combined with a $\Delta\log P$ (or $\Delta\log D$) index (log*P*-Dual filter, **Fig 6.1c**). For the latter approach, a $\Delta\log P$ (or $\Delta\log D$) value of 0.4 was used as a criterion. The log*P* (or log*D*) value of the most structurally similar ion (based on the TS score) was first compared with that of the target ion. In the case that the absolute value of $\Delta\log P$ (or $\Delta\log D$) between those two ions was greater than 0.4, $\Delta\log P$ (or $\Delta\log D$) between the second most similar ion and the target ion was checked and so forth until the selection criterion was satisfied, where the similar ion chosen in this procedure was assigned as the surrogate ion to apply the *k*-ratio filter (log*P*-Dual filter, **Fig. 6.1c**). The TS scores of

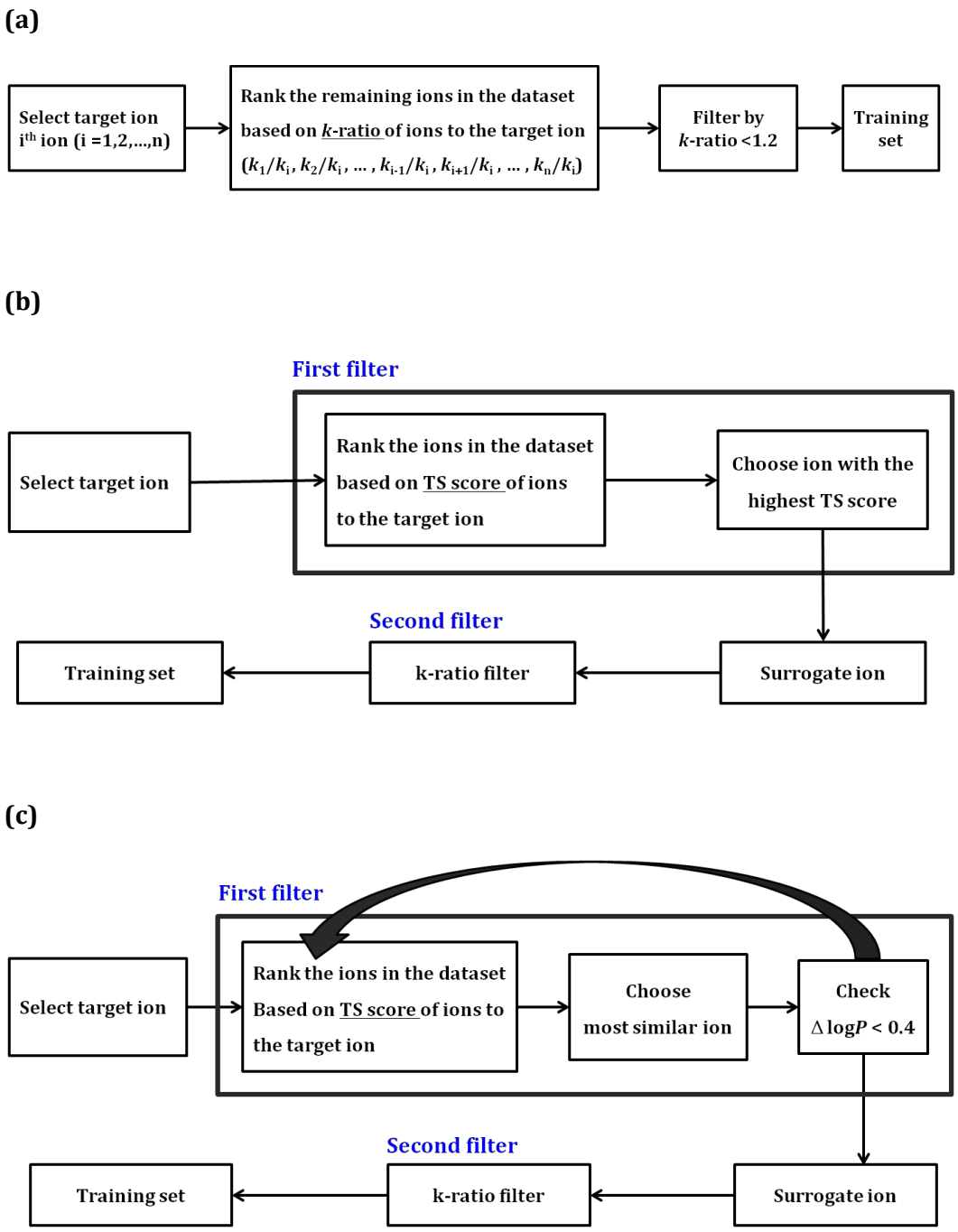


Figure 6.1 Schematic diagrams of different types of filters for similarity searching: (a) *k*-ratio filter (<1.2), (b) Dual filter (TS (>0.8) and *k*-ratio (<1.2) filter) and (c) log*P*-Dual filter (TS (>0.5), Δlog*P* (<0.4) and *k*-ratio (>1.2) filter).

ions were calculated using JChem for Excel (ChemAxon, Budapest, Hungary). Log*P* and log*D* (at pH 2) values were calculated by use of ACD/i-Lab program [37] and are presented in **Table 6.1**, along with *k* values at 10 mM MSA (used for the *k*-ratio filtering).

6.2.4. QSRR modelling by GA-PLS

QSRR models for *a*- and *b*-values were generated by PLS regression following the feature selection by a genetic algorithm (GA). QSRR modelling was carried out by automatically running a GA-PLS algorithm in Matlab R2015a (MathWorks, Natick, MA, USA). The algorithm was customised by modifying Matlab routines originally written by Leardi [38]. Prior to the GA-PLS modelling, almost constant descriptors, *i.e.*, descriptors having the same values for all but the maximum of 5 ions, were removed from the training set of each ion. The GA was performed using the following parameters. The size of original population was 50, the maximum selection probability was 20 variables/chromosome, the average selection probability was 10 variables/chromosome, cross-over probability was 50%, mutation probability was 1% and the number of evaluations prior to a backward elimination phase was 100 [15]. The resulting local models of *a*- and *b*-values, and hence the retention times, were externally validated where the same criteria for the same validation parameters (*MAE*, *RMSEP*, $Q_{ext(F2)}^2$, *s*, and (R^2 -R_0^2)/ R^2) used in Chapter 5 were applied.

6.3. Results and Discussion

6.3.1. Determination of *a*- and *b*-values for QSRR modelling

Chapter 4 has shown that QSRRs can be applied successfully to the linear solvent strength (LSS) model to predict *a*- and *b*-values and hence retention times of inorganic and low molecular-weight anions not involved in the model generation, based only on their chemical structures [39]. In this chapter, the QSRR approach, combined with the

Table 6.1 Parameters for compound clustering and eligible ions (*i.e.*, those for which a training set of >7 compounds existed) for each clustering method used.

ID	Compound	*k* at 10 mM MSA	Log*P*	Log *D* at pH 2	Eligible ions		
					*k*R filter	TS-*k*R filter	TS-ΔlogP -*k*R filter
1	1-3-Methoxyphenyl-2-methylaminoethanol	1.48	0.62	-2.6	✓	✓	✓
2	1-Butylamine	1.00	0.93	-2.36	✓		✓
3	1-Methyl-3-phenylpropylamine	2.19	2.18	-0.99	✓	✓	✓
4	2-2-Aminoethoxyethanol	0.70	-1.37	-4.25	✓		✓
5	2-Amino-1-phenylethanol	1.19	0.46	-2.89	✓		✓
6	2-Diethylaminoethanol	0.98	0.74	-2.56	✓	✓	
7	2-Dimethylaminoethanol	0.85	-0.33	-3.33	✓		✓
8	2-Ethylaminoethanol	0.77	-0.44	-3.54	✓	✓	✓
9	2-Methylaminoethanol	0.72	-0.97	-3.92	✓	✓	✓
10	3-Amino-1-propanol	0.66	-1.12	-4.75	✓		✓
11	3-Methoxytyramine	0.77	0.43	-2.75	✓	✓	✓
12	3-Methylphenethylamine	1.52	1.92	-1.32	✓		✓
13	3-Phenylpropylamine	1.99	1.83	-1.33	✓	✓	✓
14	4-Epitetracycline	1.67	-1.47	-3.85	✓	✓	✓
15	4-Phenylbutylamine	2.55	2.36	-0.81	✓		✓
16	5-Amino-1-pentanol	0.75	-0.55	-4.51	✓		
17	Acebutolol	1.53	1.95	-1.61	✓		✓
18	Alprenolol	3.14	2.88	-0.49			
19	Amino-2-propanol	0.67	-0.96	-4.49	✓		✓
20	Amoxicillin	0.56	0.92	-4.02	✓		
21	Atenolol	1.00	0.10	-3.01	✓	✓	

22	Betaxolol	3.38	2.69	-0.13		✓	
23	Bethanechol	1.16	-3.95	-3.92	✓	✓	✓
24	Bisoprolol	2.21	2.14	-0.70	✓		✓
25	Carbachol	1.11	-3.86	-4.19	✓	✓	✓
26	Carvedilol	11.20	4.11	0.72			
27	Celiprolol	2.12	2.19	-0.93	✓		✓
28	Chlortetracycline	2.93	-0.53	-3.22		✓	
29	Choline	1.05	-3.70	-3.86	✓	✓	
30	Cimetidine	0.97	0.07	-3.41	✓		
31	Clenbuterol	2.31	2.61	-0.53	✓		
32	Clomipramine	18.18	5.39	2.20			
33	Clonidine	2.18	1.41	-0.79	✓		
34	Clorprenaline	1.73	2.18	-0.91	✓		
35	Cyclohexylamine	1.28	1.39	-1.75	✓		
36	Diethanolamine	0.67	-1.50	-4.42	✓	✓	✓
37	Dimethylamine	0.80	-0.43	-3.26	✓	✓	✓
38	Diphenhydramine	5.30	3.66	0.77			
39	Dipyridamole	4.99	-1.22	-2.10			
40	Dopamine	0.59	0.12	-2.96	✓	✓	✓
41	Doxepin	7.24	3.86	1.14			
42	Doxycycline	3.03	-0.54	-3.45		✓	
43	Esmolol	1.81	1.91	-1.08	✓		✓
44	Ethanolamine	0.63	-1.31	-4.61	✓		✓
45	Etilefrine	0.76	0.50	-3.12	✓	✓	✓
46	Fenoterol	0.75	0.89	-2.16	✓	✓	✓
47	Hordenine	1.00	1.40	-1.30	✓	✓	✓
48	Hydroxyzine	9.68	2.03	-0.15			
49	Imipramine	10.15	4.80	1.66			
50	Isoprenaline	0.63	0.25	-3.46	✓	✓	✓

51	Labetalol	2.79	2.31	-0.71			
52	Metanephrine	0.70	-0.33	-3.59	✓	✓	✓
53	Metaproterenol	0.58	0.13	-3.18	✓	✓	✓
54	Methylamine	0.69	-0.66	-4.08	✓		✓
55	Metoprolol	1.66	1.79	-1.05	✓		✓
56	Mexiletine	2.03	2.16	-0.78	✓		
57	Morpholine	0.91	-1.08	-3.83	✓		✓
58	Nadolol	1.04	1.29	-2.19	✓		
59	Nebivolol	9.45	3.67	0.71			
60	Neostigmine	1.66	-3.03	-2.27	✓		
61	N-methyldiethanolamine	0.78	-0.72	-3.89	✓	✓	✓
62	N-methylphenethylamine	1.41	1.60	-1.22	✓	✓	✓
63	N-methylpyrrolidine	1.19	0.76	-2.10	✓		
64	Norepinephrine	0.48	-0.88	-4.27		✓	✓
65	Norfenefrine	0.61	-0.28	-3.80	✓	✓	✓
66	Normetanephrine	0.62	-0.57	-4.02	✓	✓	✓
67	Octopamine	0.59	-0.28	-3.8	✓	✓	✓
68	Oxprenolol	2.32	2.29	-1.09	✓		✓
69	Oxytetracycline	1.48	-1.50	-4.49	✓	✓	✓
70	Penbutolol	7.97	4.17	0.79			
71	Phenethylamine	1.21	1.46	-1.66	✓	✓	✓
72	Phenylalanine	0.70	1.11	-1.68	✓		
73	Phenylephrine	0.70	-0.03	-3.41	✓	✓	✓
74	Pindolol	1.73	1.97	-1.14	✓		
75	Promethazine	9.68	4.78	1.49			
76	Propranolol	4.05	3.10	-0.17			
77	Propylamine	0.83	0.40	-2.69	✓		✓
78	Pyrrobutamine	22.02	5.28	2.11			
79	Salbutamol	0.70	0.01	-2.91	✓		✓

80	Serotonin	0.89	0.21	-2.28	✓		
81	Sulpiride	1.14	0.45	-2.10	✓		
82	Synephrine	0.67	-0.03	-3.41	✓	✓	✓
83	tert-Butylamine	0.81	0.56	-2.51	✓		✓
84	Triethylamine	1.19	1.66	-1.34	✓		
85	Trimethylamine	0.94	0.06	-2.74	✓	✓	✓
86	Tryptophan	0.92	1.04	-1.28	✓		
87	Tyramine	0.74	0.72	-2.39	✓	✓	✓

*k*R denotes filtering by *k*-ratio only. TS-*k*R denotes dual filtering by Tanimoto Similarity and *k*-ratio. TS-Δlog*P*-*k*R denotes denotes filtering by a combined Tanimoto similarity and Δlog*P* filter followed by *k*-ratio filtering.

LSS model, was applied to 87 larger molecular-weight organic cations (molecular mass up to 506) mainly of pharmaceutical interest. The retention times of the cations were measured using isocratic eluent compositions on the CS17 column. The two retention parameters (*a*- and *b*-values) used to generate the QSRR models were estimated experimentally for each ion by plotting the logarithm of the retention factor (log*k*) against the logarithm of the eluent concentrations under five isocratic mobile phase conditions (5, 10, 20, 30 and 40 mM MSA with 36% ACN) using the following LSS model:

$$\log k = a - b\log[\text{MSA}] \qquad \textbf{(6.2)}$$

where *a* and *b* are the intercept and slope of the log*k* *vs.* log[MSA] plot, respectively, and the slope *b* relates to the charge ratio of the analyte ion to the eluent ion [21, 40]. Good correlation between log*k* *versus* log[MSA] (R^2 >0.992) (**Table 5.2**) was observed for all the cations under study, which supports the applicability of the LSS model to the QSRR study [21, 41]. The electrostatic interaction between the larger organic analytes and the stationary phase is explained well by **Eq. 6.2**, even in the presence of organic solvent (36% acetonitrile added to the eluent) [41]. The use of a fixed amount of organic solvent can lead to a constant contribution of the hydrophobic interaction to the retention in IC, regardless of variations in the eluent concentration of MSA, allowing the independent prediction of ion-exchange retention behaviour of these organic ionogenic compounds based on the LSS model [21]. However, there can still exist some hydrophobic interactions of larger organic analytes with the stationary phase, despite the addition of the organic solvent to the eluent. This was evidenced in the present study by some *b*-value deviations from the theoretical value (*i.e.*, +1 for the monovalent ions) [41]. Generally, the slopes (*i.e.*, *b*-values) were lower than the theoretical value. For example, the *b*-values of the hydrophobic analytes penbutolol (log*P* = 4.17) and pyrrobutamine (log*P* = 5.28) in our study were as low as 0.84 and 0.86, respectively. Hydrophilic alcohols and low molecular weight ions (such as methylamine [molecular

mass 32.08] and dimethylamine [molecular mass 46.11]) where there would be expected to be less hydrophobic interaction availability had *b*-values that were close to 1. The organic solvent content was determined by limitations of the suppressor and the column (max. 40% for the suppressor).

As the *b*-values in the LSS model relate to the analyte's charge, it seems reasonable to cluster the compounds based on their charges and to ensure that all modelling is performed only using ions of the same charge as the target ion. At the same time, the charges of compounds need to be constant within the range of eluent concentrations under study (5 to 40 mM MSA) to calculate their molecular descriptors. Accordingly, 87 monovalent cations were chosen from the dataset by removing both divalent ions and ions with variable charges from the initial set of 114 cations. For this procedure, the prediction of the charges of compounds was performed by considering their pK_a values (data not shown) using the ACD/iLab programme (section 2.3).

6.3.2. *k*-ratio filter-based QSRR modelling (Figure 6.1a)

Retention factor filtering (or *k*-ratio filtering) is a compound clustering method where compounds which are chromatographically similar to the target compound (based on a predetermined cut-off value of the ratio of the retention factors of the target compound and the database compound) are included in the training set to build a local QSRR model for predicting the retention of the target compound [15]. The fundamental premise of the *k*-ratio filter approach is that better prediction accuracy will be achieved by basing the modelling steps on a training set that contains only those database compounds which are chromatographically similar to the target compound, as evidenced by similar retention factors. This *k*-ratio filtering has led to the successful QSRR modelling of inorganic and small organic anions on two columns (AS20 and AS19) [15] and was further investigated in the present chapter as a method to cluster larger organic cations on the CS17 column. Specifically, QSRR modelling was performed using

training sets comprising all the chromatographically similar ions to a target ion, based on a *k*-ratio cut-off value of 1.2, and containing at least seven ions as the minimum size of the training set [20]. The *k*-ratio filtering resulted in 69 eligible ions among the 87 ions in the study, where longer retained ions, having insufficient similar ions in the training set (*i.e.*, less than seven), were not considered.

Fig. 6.2 shows strong correlations between predicted and measured *a*-, *b*- and t_R-values were obtained for *k*-ratio filtering. with *RMSEP* values of 0.08, 0.03, and 0.22 min, respectively. This implies that the generated models are very predictive and accurate. The predictive performance parameters of the developed PLS models are presented in **Table 6.2**. The predicted retention times in **Fig. 6.2c** were calculated by substituting predicted *a*- and *b*-values of QSRR models into the LSS model **(Eq. 6.2)** at the corresponding eluent concentrations (5, 10, 20, 30 and 40 mM MSA) [20, 40]. Thus, in a manner similar to that seen in the previous study on inorganic and small organic anions [15], a *k*-ratio filter-based compound clustering approach produced accurate and predictive QSRR models for larger organic cations. Despite the excellent predictive performance of the *k*-ratio filter-based QSRR modelling, this approach has a critical limitation in that the filtering method is impractical to apply to the retention prediction for new analytes for which retention data are not known [15]. Indeed, the primary goal of the QSRR procedure is to predict these retention data.

As an alternative chromatographic similarity clustering index, the log*P* ratio and log*D* ratio of ions (to the target ion) were investigated since these two indices reflect the hydrophobicity of organic compounds (as the molecular and ionic forms, respectively) and can accordingly indicate the extent of any hydrophobic contributions to the retention in IC. Moreover, estimates of these two indices can be calculated directly from the chemical structures. **Fig. 6.3** shows log*k* generally increases when log*P* and log*D* increase. This implies that hydrophobic interactions still exist for larger

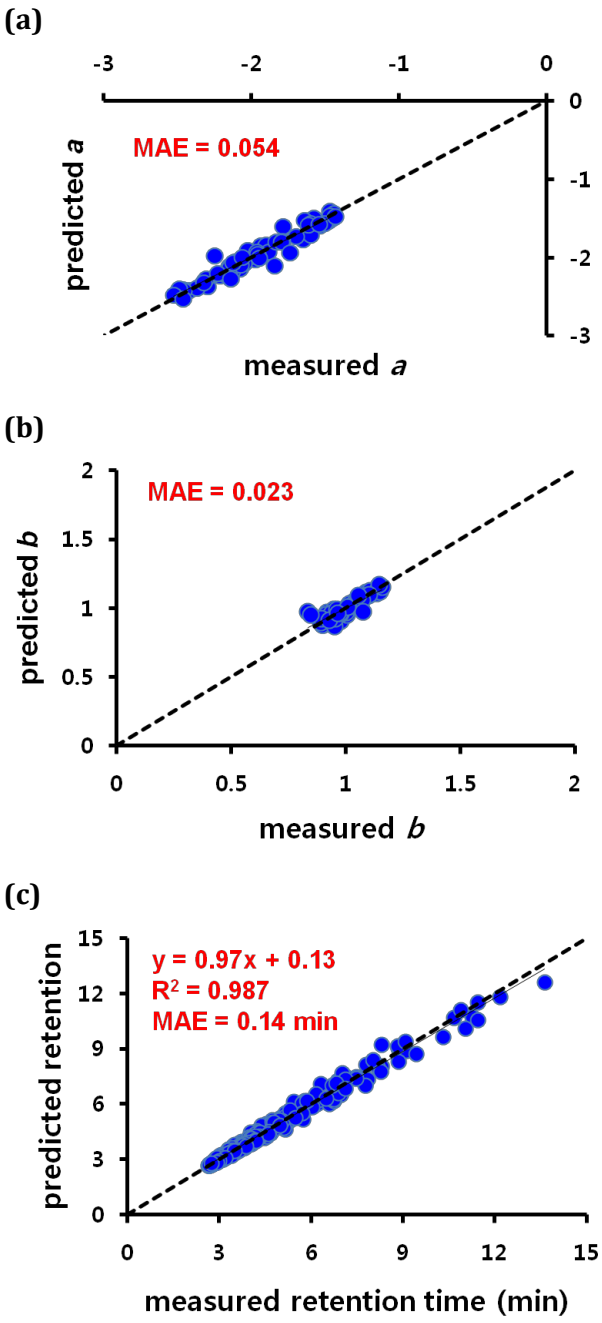
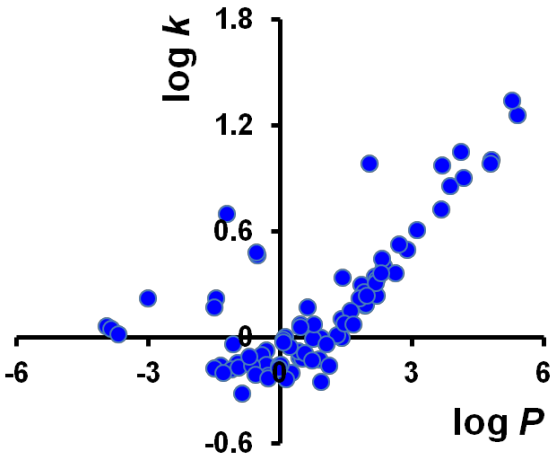


Figure 6.2 *k*-ratio (*k*-ratio < 1.2)filter-based QSRR modelling. Correlations between predicted and measured (a) *a*-, (b) *b*- and (c) t_R-values for 69 eligible organic cations on CS17 column (2 mm I.D. x 250 mm). The *k* ratios were calculated using the *k* values obtained at 10 mM MSA in the presence of 36% ACN.

(a)



(b)

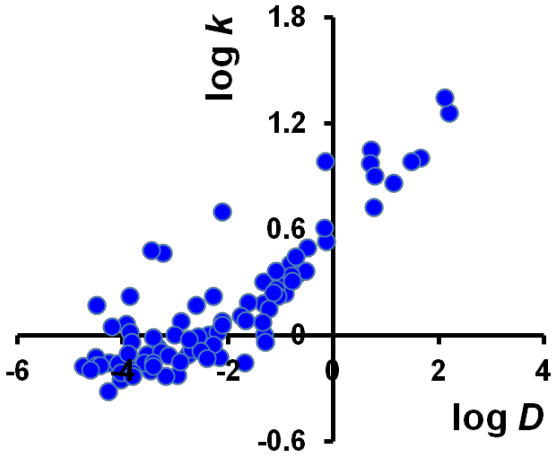


Figure 6.3 log*k vs.* (a) log*P* and (b) log*D*. The *k* values were obtained at 10 mM MSA with 36% ACN and the log*D* were calculated at pH 2.

Table 6.2 Statistical and external validation parameters. The Model name denotes filtering methods, and whether *a*- , *b*- or t_R-values are being modelled, and kR and TS denote *k*-ratio and Tanimoto similarity, respectively. *s* denotes slope through the origin.

Model	$Q^2_{ext(F2)}$	*RMSEP*	*MAE*	s_o	R^2	$\frac{(R^2 - R_0^2)}{R_0^2}$
*k*R-a	0.94	0.08	0.05	0.998	0.94	0.002
*k*R-b	0.82	0.03	0.02	1.000	0.82	0.03
*k*R-t_R	0.99	0.22	0.14	0.995	0.99	0.0009
TS-*k*R-a	0.82	0.13	0.08	1.003	0.82	0.101
TS-*k*R-b	0.83	0.03	0.02	0.999	0.84	0.02
TS-*k*R- t_R	0.75	1.07	0.44	0.899	0.77	0.13
TS-Δlog*P* -*k*R-a	0.96	0.06	0.04	1.005	0.96	0
TS-Δlog*P* -*k*R-b	0.95	0.02	0.01	1.003	0.95	0.04
TS-Δlog*P* -*k*R- t_R	0.96	0.38	0.18	0.978	0.96	0.003
TS-Δlog*D* -*k*R-a	0.93	0.08	0.06	1.002	0.93	0.002
TS-Δlog*D* -*k*R-b	0.91	0.02	0.02	1.002	0.92	0.009
TS-Δlog*D* -*k*R- t_R	0.92	0.55	0.27	0.987	0.92	0.003

*k*R denotes filtering by *k*-ratio only. TS-*k*R denotes dual filtering by Tanimoto similarity and *k*-ratio. TS-Δlog*P*-*k*R denotes filtering by a combined Tanimoto similarity and Δlog*P* filter followed by *k*-ratio filtering. TS-Δ*logD*-*k*R denotes filtering by a combined Tanimoto similarity and logD filter followed by k-ratio filtering.

organic cations although this interaction was reduced by adding 36% acetonitrile. However, both log*P*- and log*D*-ratio filter-based QSRR modelling (using clustering based on cut-off values of 1.4 and 1.2 for log*P* and log*D*, respectively) were unsuccessful in the external validation of models for t_R-values (graphs not shown). The log*P*-ratio filter-based model failed to meet two of the validation criteria (*MAE* of 1.2 min and slope *s* of 0.82), and the log*D*-ratio approach failed to meet three validation criteria ($Q_{ext(F2)}^2$ of 0.6, R^2 of 0.59, and $(R^2 - R_0^2)/R^2$ of 0.35) [8, 15, 42]. This result is also supported by the deviation from the 45 degree line in **Figure 6.3** showing that while hydrophobicity is important, the IC retention mechanism is mainly affected by electrostatic interactions, as well as other factors such as the polarisability of the ions, but under the conditions used, analyte hydrophobicity is not a good predictor of retention.

6.3.3. Dual filter-based QSRR modelling

A dual-filter strategy **(Fig 6.1b)** proposed and examined on HILIC and RPLC datasets [15] was further investigated on the present IC dataset consisting of larger organic cations, to address the main limitation of *k*-ratio filtering, which is that it is impractical to use for retention prediction of unknown compounds. Briefly, the most similar ion to the target ion (labelled as the "surrogate ion") was first selected using the Tanimoto similarity (TS) score (labelled as the first filter) and the *k*-ratio of ions in the database to the surrogate ion (labelled as the second filter) was then calculated to identify a training set of analytes having *k*-ratio values <1.2 **(Fig. 6.1b)**. The underlying premise is that the TS score may adequately reflect chromatographic similarity in IC due to the relatively simple retention mechanism which applies [15]. Additionally, Chapter 5 showed that predictive and accurate QSRR models available for IC "scoping" method development can be generated in the case where an appropriate level of the TS threshold value (TS >0.6) is applied to select a cluster of compounds for the training set

[20]. **Figure 6.4** supports the premise mentioned above, showing the difference in retention factor between any given ion and its most similar ion for all the 87 cations plotted against their pair-wise TS score. The difference in retention factor (Δk) generally decreased with increasing pair-wise TS scores, implying that the most similar ion identified by TS score can exhibit similar k values to the target ion selected for a local QSRR model. Accordingly, k-ratio filtering based on the surrogate ion (*i.e.*, the most similar ion by TS to the target ion) will be likely to identify a training set which has close chromatographic similarity to the target ion. **Figure 6.5** illustrates correlations between predicted and measured a-, b- and t_R- values for 37 eligible ions, based on the dual filter. Among a total of 87 ions, 38 ions (having TS >0.8) were considered for application of the dual filter approach **(Fig. 6.4)**. When the second filter (k-ratio filter) was applied to the most similar ion to the unknown, one ion (ID55, **Table 6.1**) had only 4 similar ions with a k-ratio <1.2 in the training set and hence was removed (because the minimum number needed in the training set for the modelling was 7) **(Table 6.1)**. Although predicted a- and b- values generally agreed with the corresponding measured values **(Fig. 6.5a** and **6.5b)**, six ions (ID 1, 21, 22, 28, 42, 61 in **Table 6.1**) appeared to be deviating more radically from the 45 degree line in the correlation plot for a-values **(Fig. 6.5a)**, which resulted in severe errors in retention time predictions (mean absolute errors (*MAE*) >2.5 min) **(Fig. 6.5c)**. These data points (except one data point (ID 61, Δk value = 0.2) have relatively large Δk values (Δk values >0.65). Considering the difference in log$P$ between the surrogate ion and the target ion for these outliers (Δlog$P$ >0.89), the errors in prediction may possibly be due to the significant difference in hydrophobicity of these compounds and the target ion. For example, compound ID1 is hydrophobic (log$P$ of 0.62, **Table 6.1**) whereas compound ID52 is hydrophilic (log$P$ of −0.33, **Table 6.1**). Despite their structural similarity (TS of 0.87), the relatively high difference in their hydrophobicity (Δlog$P$ of 0.95) led to a

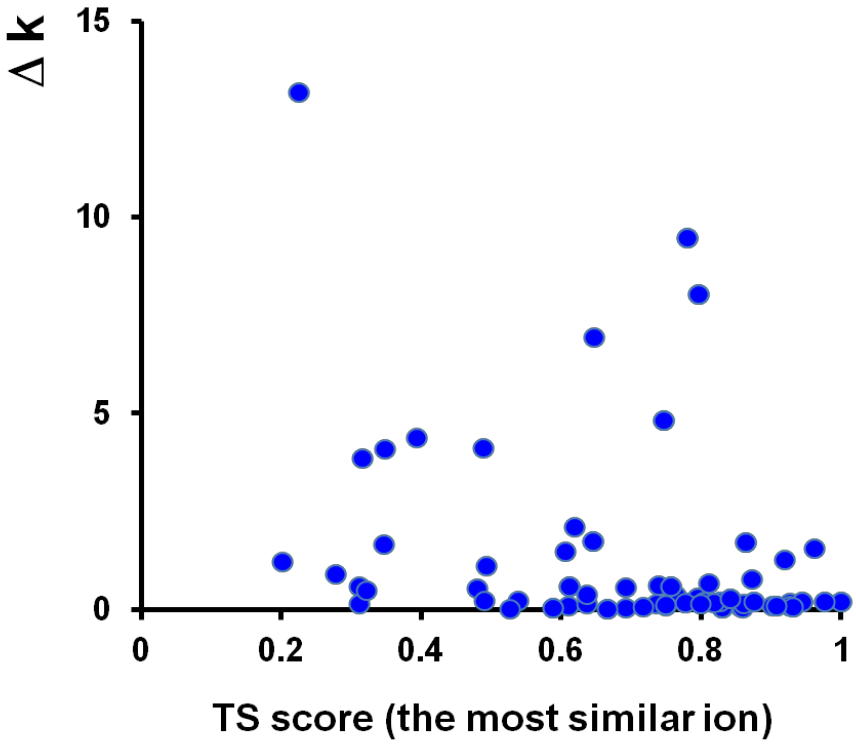


Figure 6.4 Δ*k vs.* Tanimoto similarity (TS) scores between the most similar ion and the test ion.

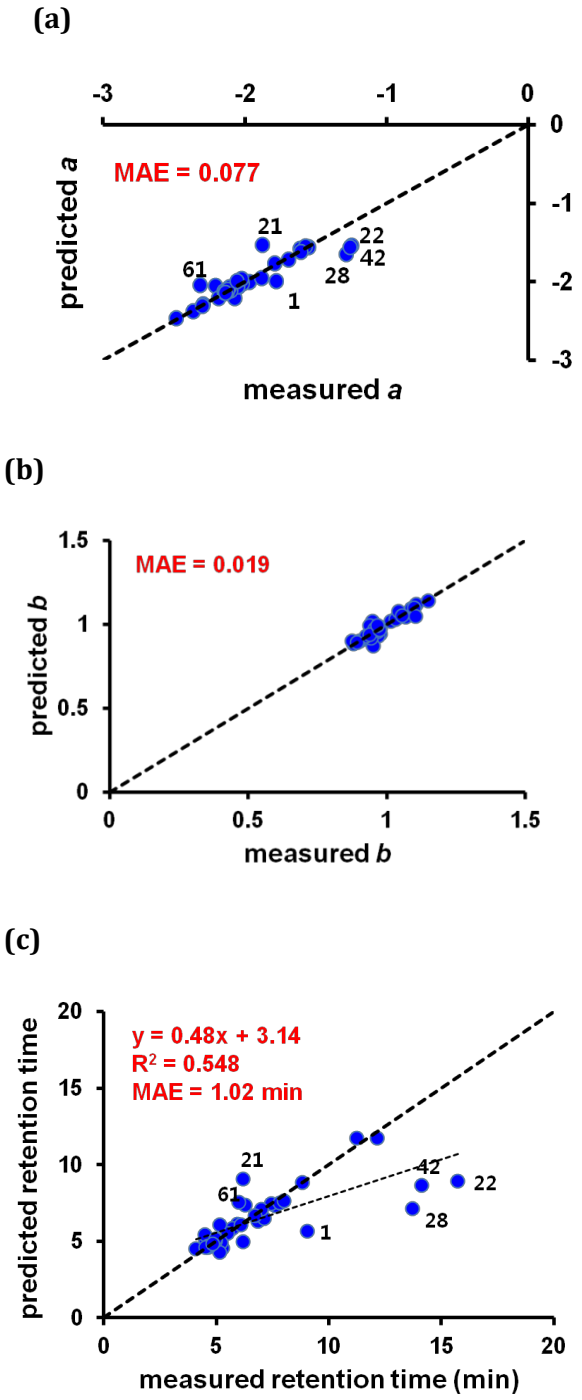


Figure 6.5 TS (>0.8), combined with *k*-ratio (*k*-ratio <1.2), filter-based QSRR modelling. Correlations between predicted and measured (a) *a*-, (b) *b*- and (c) t_R-values for 37 eligible organic cations on CS17 column (2 mm I.D. x 250 mm). The t_R-values at 5 mM MSA were used. The *k* ratios were calculated using the *k* values obtained at 10 mM MSA with 36% ACN.

relatively large difference in retention (Δk of 0.78). This implies that the extent of hydrophobicity of compounds (*i.e.*, $\Delta \log P$ between the most similar ion and the target ion) also needs to be taken into account when clustering compounds for the modelling. Consequently, the models for *a*-values using the dual filter approach failed external validation for two criteria (R^2 of 0.82 and $(R^2 - R_0^2)/ R^2$ of 0.101), and hence models for t_R-values were also not successful for these two criteria (R^2 of 0.77 and $(R^2 - R_0^2)/R^2$ of 0.13) **(Table 6.2)**.

Therefore, to better identify the most chromatographically similar ion to the target ion, the dual-filter approach was modified by adding an intermediate step comparing the log*P* value of the surrogate ion with that of the unknown (log*P*-Dual filter, **Fig 6.1.(c)**). Accordingly, for any given target ion, the ions in the database were firstly ranked based on their TS scores to that ion. Then, the highest ranked ion in the database for which $\Delta \log P$ (compared to the target ion) was less than 0.4 was selected as the surrogate ion for the subsequent *k*-ratio filtering **(Fig. 6.1c)**, followed by QSRR modelling. This log*P*-Dual filter procedure was applied only for the ions having TS >0.5. **Figure 6.6** shows correlations between predicted and measured *a*-, *b*- and t_R-values for the 50 eligible ions based on the log*P*-Dual filter approach. Accurate and predictive QSRR models for *a*- and *b*-values resulted in good agreement between predicted and measured retention times (for a total of 250 data points covering 5 eluent concentrations), with an R^2-value of 0.96 and *RMSEP* of 0.38 min **(Fig. 6.6c)**. Worth noting is that the prediction error in retention time of an outlier ID 61 **(Table 6.1**, **Fig. 6.5)** was significantly improved (*MAE* of 0.05 min) using the second most similar ion ID 7 (TS 0.83) instead of the most similar ion ID6 (TS 0.94) (*MAE* of 1.5 min), as the $\Delta \log P$ value between ID 7 and ID 61 was 0.39, far less than that between ID 6 and ID 61 **(**1.46, **Table 6.1)**. Since the analytes under study are organic cationic analytes and the hydrophobicity of ionic species is generally represented by log*D*, a modified dual-

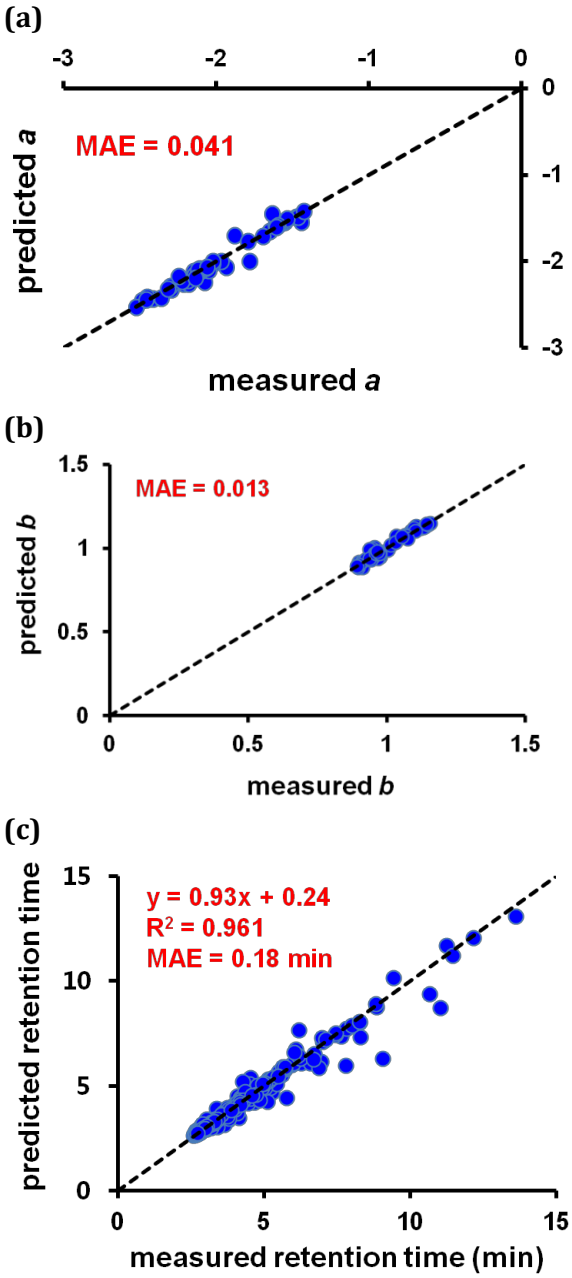


Figure 6.6 *k*-ratio (*k*-ratio <1.2), following TS (>0.5) combined with Δlog*P* (<0.4) filter-based QSRR modelling. Correlations between predicted and measured (a) *a*-, (b) *b*- and (c) t_R-values for 50 eligible organic cations on CS17 column (2 mm I.D. x 250 mm). The *k* ratios were calculated using the *k* values obtained at 10 mM MSA with 36% ACN.

filtering method employing log*D* instead of log*P* was also investigated to check whether the prediction of retention times of these organic cations could be improved. Interestingly, compared to log*P*, the error in retention time prediction was slightly higher (*RMSEP* of 0.55 min), which is supported by poorer correlation between predicted and measured retention time **(Fig. 6.7c)**. This is probably because log*D* values change according to pH. In this study, log*D* values calculated at pH 2 were used to check Δlog*D*. The change in log*D* with pH can therefore cause inaccuracy in the QSRR models, especially since the actual eluent pH may further change due to the addition of organic modifier. The modified dual-filter using log*P* provides more robust and predictive models. Additionally, the modified dual filtering using log*P* or log*D* provided a larger number of eligible ions (50 ions) than the TS filter (23 ions) [20]. This is because by introducing the Δlog*P* measure, ions based on lower TS threshold values (a change from 0.8 **(Fig. 6.4)** to 0.5) could be explored to search for the chromatographically most similar ions to the target. The characteristics of the developed models based on the two dual filtering approaches (TS-Δlog*P*- *k* ratio filter and TS-Δlog*D*- *k* ratio filter) are summarised in **Table 6.2**.

6.4. Conclusions

In this chapter, localised QSRR modelling for two retention parameters (*a*- and *b*-values) in the LSS model in ion chromatography was successfully demonstrated for 50 organic cations of mainly pharmaceutical interest on a Thermo Fisher Scientific CS 17 column, by employing several chromatographic similarity-based compound clustering methods. The *k*-ratio filtering approach of selecting a training set based on ions having *k*-ratio <1.2 provided the best QSRR models for 69 eligible cations, providing excellent predictive power for retention time prediction ($Q_{ext(F2)}^2$ of 0.99 and *RMSEP* of 0.22 min). To address the inherent impracticality of *k*-ratio filtering which does not cover the

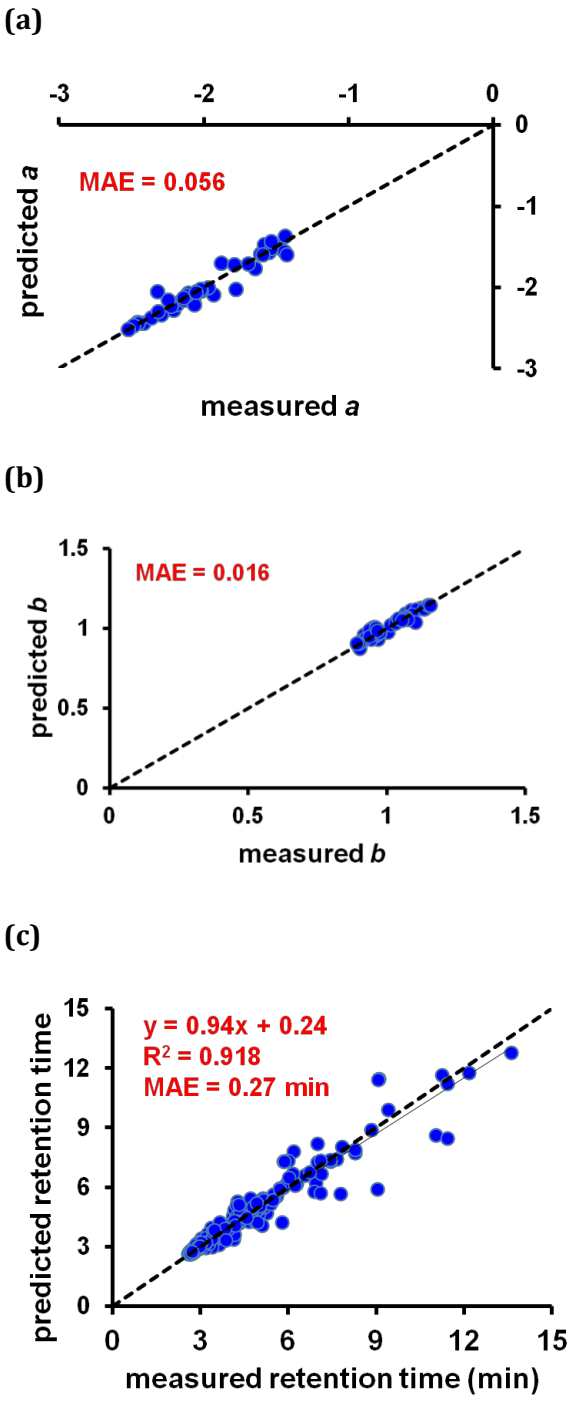


Figure 6.7 *k*-ratio (*k*-ratio <1.2), following TS (>0.5) combined with Δlog*D* (<0.4), filter-based QSRR modelling. Correlations between predicted and measured (a) *a*-, (b) *b*- and (c) t_R-values for 48 eligible organic cations on CS17 column (2 mm I.D. x 250 mm). The *k* ratios were calculated using the *k* values obtained at 10 mM MSA with 36% ACN and log*D* values were obtained at pH 2.

retention prediction for unknown compounds, a dual filtering method was further investigated by combining an initial Tanimoto similarity (TS) filter with a secondary *k*-ratio (<1.2) filter. The concept of this approach is to identify from the dataset the most structurally similar ion (compared to the unknown, TS >0.8) that adequately reflects the IC retention and then to use this surrogate ion to select a training set for QSSR modelling that comprises only those ions having chromatographicsimilarity (*k*-ratio <1.2). While the hydrophobic interaction between analytes and the stationary phase was reduced by the addition of organic modifier (36% acetonitrile) to the eluent, the introduction of a hydrophobic property (Δlog*P*) to select a training set having the most chromatographically similar ions can substantially improve the prediction accuracy of the derived models. The introduction of the Δlog*P* measure can also lower the threshold value of the TS score (TS >0.5), resulting in more eligible ions (50 ions) compared to 37 ions using the dual filter without the Δlog*P* measure. Thus, a promising dual filtering approach (log*P*-Dual filter) was developed, producing successful QSRR models with good accuracy and predictive power ($Q_{ext(F2)}^2$ of 0.96 and *RMSEP* of 0.38 min), in which a surrogate ion for the target ion was first selected considering both Tanimoto (TS >0.5) and Δlog*P* similarity indicies (Δlog*P* <0.4) and the *k*-ratio (<1.2) filter was then applied using the surrogate ion instead of the target ion.

The proposed approach shows great promise for successful localised QSRR modelling for *a*- and *b*-values in the LSS model and hence retention times of target ions under a broad range of eluent conditions, leading to robust and rapid scoping method development in IC. Future work may include expanding the proposed approach to larger organic anions as well as to a wider range of columns.

6.5. References

[1] R. Kaliszan, M.A.V. Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure–retention relationships, J. Chromatogr. A 855 (1999) 455–486.

[2] E. Daghir-Wojtkowiak, S. Studzińska, B. Buszewski, R. Kaliszan, M.J. Markuszewski, Quantitative structure–retention relationships of ionic liquid cations in characterization of stationary phases for HPLC, Anal. Methods 6 (2014) 1189.

[3] A. Plenis, L. Konieczna, N. Miekus, T. Baczek, Development of the HPLC Method for Simultaneous Determination of Lidocaine Hydrochloride and Tribenoside Along with Their Impurities Supported by the QSRR Approach, Chromatographia, 76 (2013) 255-265.

[4] N. Kritikos, A. Tsantili-Kakoulidou, Y.L. Loukas, Y. Dotsikas, Liquid chromatography coupled to quadrupole-time of flight tandem mass spectrometry based quantitative structure-retention relationships of amino acid analogues derivatized via n-propyl chloroformate mediated reaction, J. Chromatogr. A 1403 (2015) 70-80.

[5] P.E. Morgan, D.J. Barlow, M. Hanna-Brown, R.J. Flanagan, Artificial Neural Network Modelling of the Retention of Acidic Analytes in Strong Anion-Exchange HPLC: Elucidation of Structure-Retention Relationships, Chromatographia, 75 (2012) 693-700.

[6] M. Goodarzi, R. Jensen, Y. Vander Heyden, QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions, J. Chromatogr. B 910 (2012) 84-94.

[7] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, Classification and regression tree analysis for molecular descriptor selection and

retention prediction in chromatographic quantitative structure–retention relationship studies, J. Chromatogr. A 988 (2003) 261-276.

[8] M. Talebi, G. Schuster, R.A. Shellie, R. Szucs, P.R. Haddad, Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography, J. Chromatogra. A 1424 (2015) 69-76.

[9] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemom. Intell. Lab. Syst. 76 (2005) 185-196.

[10] G. Carlucci, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure-retention relationships, Anal. Chim. Acta 601 (2007) 68-76.

[11] Š. Ukić, M. Novak, P. Žuvela, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part I: Conventional QSRR approach, Chromatographia, 77 (2014) 985-996.

[12] Š. Ukić, M. Novak, A. Vlahović, N. Avdalović, Y. Liu, B. Buszewski, T. Bolanča, Development of gradient retention model in ion chromatography. Part II: Artificial intelligence QSRR approach, Chromatographia, 77 (2014) 997-1007.

[13] T. Baczek, R. Kaliszan, Combination of linear solvent strength model and quantitative structure–retention relationships as a comprehensive procedure of approximate prediction of retention in gradient liquid chromatography, J. Chromatogr. A 962 (2002) 41–55.

[14] K. Gorynski, B. Bojko, A. Nowaczyk, A. Bucinski, J. Pawliszyn, R. Kaliszan, Quantitative structure-retention relationships models for prediction of high

performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds, Anal. Chim. Acta 797 (2013) 13-19.

[15] E. Tyteca, M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad, Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Models for retention prediction: use of retention factor ratio, J. Chromatogr. A 1486 (2016) 50-58.

[16] K. Heberger, Quantitative structure-(chromatographic) retention relationships, J. Chromatogr. A 1158 (2007) 273-305.

[17] K. Muteki, J.E. Morgado, G.L. Reid, J. Wang, G. Xue, F.W. Riley, J.W. Harwood, D.T. Fortin, I.J. Miller, Quantitative structure retention relationship models in an analytical quality by design framework: simultaneously accounting for compound properties, mobile-phase conditions, and stationary-phase properties, Ind. Eng. Chem. Res. 52 (2013) 12269-12284.

[18] C. Wang, M.J. Skibic, R.E. Higgs, I.A. Watson, H. Bui, J. Wang, J.M. Cintron, Evaluating the performances of quantitative structure-retention relationship models with different sets of molecular descriptors and databases for high-performance liquid chromatography predictions, J. Chromatogr. A 1216 (2009) 5030-5038.

[19] M. Talebi, S.H. Park, M. Taraji, Y. Wen, R.I.J. Amos, P.R. Haddad, R.A. Shellie, R. Szucs, C.A. Pohl, J.W. Dolan, Retention time prediction based on molecular structure in pharmaceutical method development: A perspective, LCGC North America 34 (2016) 550-558.

[20] S.H. Park, M. Talebi, R.I.J. Amos, E. Tyteca, P.R. Haddad, R. Szucs, C.A. Pohl, J.W. Dolan, Towards a chromatographic similarity index to establish localised Quantitative Structure-Retention Relationships for retention prediction. II

Comparison of similarity measures for predicting retention in ion chromatography J. Chromatogr. A, in press. DOI: http://doi.org/10.1016/j.chroma.2017.02.054

[21] P. Zakaria, G.W. Dicinoski, B.K. Ng, R.A. Shellie, M. Hanna-Brown, P.R. Haddad, Application of retention modelling to the simulation of separation of organic anions in suppressed ion chromatography, J. Chromatogr. A 1216 (2009) 6600-6610.

[22] L. Escuder-Gilabert, S. Sagrado, R.M. Villanueva-Camanas, M.J. Medina-Hernandez, Quantitative structure-retention relationships for ionic and non-ionic compounds in biopartitioning micellar chromatography, Biomed. Chromatogr. 19 (2005) 155-168.

[23] M.A. Al-Haj, R. Kaliszan, B. Buszewski, Quantitative Structure-Retention Relationships with model analytes as a mean of an objective evaluation of chromatography columns, J. Chromatogr. Sci. 39 (2001) 29-38.

[24] J. Ghasemi, S. Saaidpour, QSRR prediction of the Chromatographic Retention Behavior of Painkiller Drugs, J. Chromatogr. Sci. 47 (2009) 156-163.

[25] S. Studzinska, M. Molikova, P. Kosobucki, P. Jandera, B. Buszewski, Study of the Interactions of Ionic Liquids in IC by QSRR, Chromatographia, 73 (2011) 35-44.

[26] E. Rutkowska, K. Pajak, K. Jozwiak, Lipophilicity-methods of determination and its role in medicinal chemistry, Acta Pol. Pharm. 70 (2013) 3-18.

[27] MarvinSketch, ChemAxon 2016, chemaxon.com [Accessed: April 2016].

[28] T.A. Halgren, Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94, J. Comput. Chem. 17 (1996) 490-519.

[29] T.A. Halgren, Merck molecular force field .2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions, J. Comput. Chem. 17 (1996) 520-552.

[30] T.A. Halgren, Merck molecular force field .3. Molecular geometries and vibrational frequencies for MMFF94, J. Comput. Chem. 17 (1996) 553-586.

[31] T.A. Halgren, R.B. Nachbar, Merck molecular force field .4. Conformational energies and geometries for MMFF94, J. Comput. Chem. 17 (1996) 587-615.

[32] M.J. Vainio, M.S. Johnson, Generating conformer ensembles using a multiobjective genetic algorithm, J. Chem. Inf. Model. 47 (2007) 2462-2474.

[33] J.S. Puranen, M.J. Vainio, M.S. Johnson, Accurate conformation-dependent molecular electrostatic potentials for high-throughput *in silico* drug discovery, J. Comput. Chem. 31 (2010) 1722-1732.

[34] J.J. Stewart, Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, J. Mol. Model. 19 (2013) 1-32.

[35] MOPAC 2012, Stewart Computational Chemistry, in, Colorado Springs: CO, USA, OpenMOPAC.net. [Accessed: January 2015].

[36] Dragon 6.0, Talete, Milano, Italy, 2014, talete.mi.it. [Accessed: April 2016].

[37] ACD/i-Lab Freeware, Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2016, ilab.acdlabs.com. [Accessed: December 2016].

[38] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.

[39] S.H. Park, P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, C.A. Pohl, Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model, J. Chromatogr. A 1486 (2017) 68-75.

[40] P.R. Haddad, P.E. Jackson, Ion chromatography: principles and applications, in: Journal of Chromatography Library, Elsevier, Amsterdam, The Netherlands, 1990.

[41] P. Zakaria, G. Dicinoski, M. Hanna-Brown, P.R. Haddad, Prediction of the effects of methanol and competing ion concentration on retention in the ion chromatographic separation of anionic and cationic pharmaceutically related compounds, J. Chromatogr. A 1217 (2010) 6069-6076.

[42] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of Being Earnest: Validation in the absolute essential for successful application and interpretation of QSPR models, QSAR & Combinatorial Science 22 (2003) 69-77.

Chapter 7

General conclusions

The construction of Quantitative Structure-Retention Relationships (QSRRs) has been investigated for the *in-silico* prediction of ion-exchange chromatographic retention, with an aim of not only accelerating the scoping phase of method development but also improving the robustness and ruggedness of the developed method. Localised QSRR modelling – using a training set comprised of a cluster of similar ions to the target ion – is a promising approach, and has minimised errors in retention time prediction.

Initially the retention data of anions embedded in the Virtual Column software, the only available commercial optimisation tool for IC separations, was updated by a new "porting" methodology, to allow the provision of reliable and durable input data (or the response variables) in QSRR modelling. This "porting" method had been previously developed to calibrate extensive retention databases embedded in the Virtual Column software on newly produced columns, by reflecting changes in column behaviour due to batch-to-batch variability in the manufacturing process, the production of new column versions, and column age. Since the previous porting method incurred errors on some columns, it was modified by increasing the number of representative anions used in deriving porting equations from two (chloride and thiosulfate) to six (chloride, bromide, iodide, perchlorate, sulfate, and thiosulfate). The use of a range of ions with different characteristics allowed precise reflection of the changes in column behaviour, leading to an improvement of the porting accuracy on a wide range of columns. The use of ported retention times in Virtual Column resulted in predictive errors less than 1.3% on three new columns tested (AS20, AS19, and AS11HC), providing accurate anion data that are applicable to QSRR modelling. More reliable simulations and *in silico*

optimization by the Virtual Column software can be further expected, provided that the new porting procedure is applied on all the extensive embedded databases.

Next, QSRR models using the abovementioned ported retention data of inorganic and small organic ions were generated on the three columns (49, 41 and 40 ions on AS20, AS19, and AS11HC, respectively). Instead of directly predicting retention times of analytes, two retention parameters (*a* and *b*) in the linear solvent strength (LSS) model, a well-known and accurate isocratic retention model in IC, were predicted by deriving mathematical relationships (*i.e.*, QSRR) between these two parameters and their relevant molecular descriptors. A great advantage of this approach was that retention times of the analytes can be predicted, using only the predicted *a*- and *b*-values, for all eluent compositions under isocratic and gradient eluent modes, as well as multi-step eluent profiles comprised of a combination of isocratic and linear gradient eluent modes.

QSRR models for the *a*- and *b*-values were created by multiple linear regression (MLR) using the optimal subset of molecular descriptors selected. The optimal subset of molecular descriptors was chosen by two procedures: the selection of the optimal number of molecular descriptors by an evolution algorithm (EA) and the elimination of multi-collinear descriptors. The optimal number of descriptors was determined where the highest R^2 (>0.99)-and the smallest RMSE values were obtained by the EA and the variance inflation factor (VIF) and correlation coefficient (r) in the correlation matrix were used as measures of multi-collinearity (VIF <10 and r >0.5). The validity of the generated six QSRR models was confirmed by external validation using external test sets (consisting of 10% of the total data set), resulting in acceptable predictive ability for all six models (models for *a*- and *b*-values on each of the three columns) showing ($Q_{ext(F3)}^2$ >0.7 and *RMSEP* <0.4). As a result, a predictive error (*RMSEP*) in the retention times of 1.18 min was obtained for the external test anions, implying an acceptable

level of retention time prediction for unknown inorganic and small anions over a broad range of eluent concentrations on a wide range of columns. The proposed approach (*i.e.*, QSRR modelling predicting the *a*-and *b*-values in the LSS model) can be applied to any type of column and eluent provided that suitable isocratic retention data can be collected.

Since the construction of reliable QSRR models enabling the retention prediction for target compounds is one major point of interest in QSRR modelling, localised QSRR modelling in IC has been investigated by introducing various similarity measures to improve the predictive power of the models for the two retention parameters (*a* and *b*) of the target analytes in IC. A local QSRR model for each target ion was generated using only similar ions to the target ion, clustered by the similarity searching of the database. A genetic algorithm-partial least squares (GA-PLS) regression was used due to the ability of PLS to favourably handle collinear descriptors. Two important factors of structural similarity (between ions to be modelled and target ions to be predicted) played an important role in the QSRR modelling in IC: the average structural similarity score of the ions in the training set compared to the target ion, and the number of similar ions in the training set compared to the target ion. The measure of structural similarity utilised was a pair-wise Tanimoto similarity (TS). The TS is the most popular 2D fingerprint-based similarity measure and varies from 0 to 1, with 1 indicating 100% similarity. Predictive QSRR models for target ions not included in the model creation could be generated ($Q_{ext(F2)}^2$ >0.8 and *RMSEP* <0.1) based on a TS threshold of 0.6, resulting in accurate predictions of retention times of unknown ions (*RMSEP* of 0.44 min) for broad eluent compositions. For this approach, a dataset on the CS17 column consisting of 91 cations of mainly pharmaceutical interest was employed and only 23 ions were eligible for the modelling, implying the need for larger datasets consisting of clusters of ions with high structural similarity.

A variety of chromatographic similarity indices such as the retention factor ratio (*k*-ratio) of ions to the target ion and a combination of TS and the *k*-ratio, were investigated as alternative similarity measures to increase the number of eligible ions for local modelling. The underlying premise of *k*-ratio similarity was that chromatographically similar ions, with *k*-ratio values close to 1, will be eluted in the vicinity of the target ion. When applying a *k*-ratio cut-off value of 1.2 to include chromatographically similar ions in the training set, 69 cations were eligible for the local QSRR modelling and retention time predictions of these cations were excellent ($Q_{ext(F2)}^2$ of 0.99 and *RMSEP* of 0.22 min).

Since the *k*-ratio-based clustering (or *k*-ratio filter) method cannot be used in predicting retention times of real target ions due to the lack of *a priori* knowledge of their retention, a dual filter-based clustering method, combining a TS (first) filter with a *k*-ratio (second) filter, was investigated to develop localised QSRR models enabling the retention prediction for real target ions. The most structurally similar ion to the target ion, adequately reflecting the IC retention, was identified from the dataset, among ions having TS values greater than 0.8, and was employed as a surrogate ion of the target ion to be predicted (first filter). This surrogate ion was then used to select a cluster of only chromatographically similar ions having *k*-ratio values less than 1.2 to the surrogate into a training set for QSRR modelling (second filter).

Some outliers from the 45 degree line in the plot of predicted *a* and *b* values *vs.* measured values were investigated and there were relatively large differences in hydrophobicity (log*P*) between the surrogate ion and the target. Thus, the dual filter was modified by introducing a Δlog*P* measure following the TS filter, to find a more suitable surrogate ion reflecting IC retention more accurately. The modified dual filter (*i.e.*, log*P*-Dual filter) led to successful QSRR modelling, providing good accuracy and predictive power of the generated models ($Q_{ext(F2)}^2$ of 0.96 and *RMSEP* of 0.38 min). In

addition, the proposed log*P*-Dual filter allowed more eligible ions (50 ions) for the QSRR modelling due to the lowered TS threshold value (TS >0.5) as the available criterion, in comparison to the TS filter (22 ions) and the dual filter without the Δlog*P* measure (37 ions). Future work will include optimising the threshold criteria (TS, Δlog*P* and *k*-ratio) for improved accuracy of QSRR models, and expanding the proposed filtering strategy to larger organic anions as well as to a larger variety of columns.

In conclusion, this thesis has successfully demonstrated promising compound clustering-based localised QSRR modelling in IC. The successful QSRR modelling for the two retention parameters (*a* and *b*) in the LSS model, using the log*P*-dual filter-based clustering method, has allowed an acceptable level of retention time predictions of unknown ions for broad eluent compositions, showing its great potential for rapid scoping method development in IC. Future work may include discovering the physicochemical meanings of molecular descriptors in the developed QSRR models, allowing the interpretation of IC retention mechanism (or separation properties). In addition, the characterisation and classification of various IC columns, based on the retention mechanism, may be followed, which in turn can facilitate the selection of IC column (*i.e.*, rapid scoping method development).